Devising a Toolkit to Evaluate the High Quality Endoscopy Trainer

MD Dissertation

Institute of Health and Society

School of Medical Sciences Education Development

Louise Macdougall

September 2013

Address

Education Centre

North Tyneside General Hospital

Rake Lane

North Shields

NE29 8NH

lmacdougall@doctors.org.uk

Supervisors

Professor Roger Barton,

Dr Sally Corbett, Health Psychologist

Dr Mark Welfare, Consultant Gastroenterologist

Dr Christopher Wells, Consultant Gastroenterologist

Date of first registration: September 2010

Date of first assessment: June 2011

Date of second assessment: August 2012

Details of research degree programme undertaken: Doctor of medicine, medical education development (full time)

“It goes without saying that no man can teach successfully who is not at the same time a student”

William Osler 1905

Dedicated to my husband, Billy, thank you.

This study was granted ethical approval; it was applied for in two phases. Phase one (Cognitive interviews and Delphi process) is covered under REC 10/H904/58 and phase two (trialling the tool) was covered by REC 11/EE/0315

Abstract

Training of future endoscopists within the UK has shown to be of variable quality. Those learning endoscopy have the opportunity to attend short courses but much of their training occurs within base hospitals around the UK; this training tends to occur via an apprenticeship style model on a one-to-one basis with an experienced endoscopist, the 'trainer'. These trainers have the opportunity, and are encouraged, to attend 'Training the trainer' courses but then receive no ongoing validated feedback about their training.

Evidence suggests that gaining feedback on teaching performance can improve subsequent performance; therefore this project aimed to create a validated feedback tool which could be completed by trainees, peers or trainers as a self-reflection exercise. This tool can then be used to give formative feedback to endoscopy trainers.

Methodology

In order to create an evaluation tool a previous list of attributes that describe a high quality endoscopy trainer had already been developed (Wells 2010). This list was used to form the basis of a toolkit. The evaluation toolkit consisted of two components,

- the DOTS (Direct Observation of Teaching Skills) which could be completed after a single procedure or endoscopy session by a trainee, peer or the trainer as a self-evaluation

-the LETS (Long-term Evaluation of Teaching Skills) which could be completed at the end of a rotation either by a trainee or a trainer

Before developing the toolkit clarity of the attributes was ensured using a cognitive interviewing process. Both trainers and junior doctors were interviewed to try and ensure that the attributes were correctly interpreted by all. A total of six interviews were conducted.

In order to decide which attributes to include within the toolkit and to allocate the attributes to the two components of the toolkit a Delphi process was used. A Delphi process is a group consensus technique that can be used to gain the opinion of experts on a topic. It does not require participants to meet and is conducted via a series of

questionnaires that allow each individual to express their views and have the opportunity to revise their opinions in light of others. A two round Delphi was conducted, 'experts' were invited from four groups; expert trainers who taught on the teaching courses, base hospital trainers, trainees and nurse endoscopists.

The panel were given the opportunity to rate each attribute's suitability for the DOTS and LETS on a five point Likert scale and comment on the items. Items that gained greater than 77% consensus (i.e. were scored 4 or 5) were allocated to the toolkit. If there was a significant difference between the scores for the DOTS and the LETS they were allocated to the component of the toolkit in which they had scored highest; inconclusive items were reviewed by the panel in round two.

The Delphi process resulted in a provisional DOTS and LETS; these then required trialling to gain a further assessment of their psychometric properties. This was carried out by gaining peer evaluations of the trainer on the DOTS using the JAG 'Training the trainer' courses; trainee and trainer evaluations for both the DOTS and the LETS were collected in local units in the North East. As well as completing the toolkit trainees and trainers were also asked to complete the CTEI (Copeland and Hewson 2000), which is a validated tool for the evaluation of clinical teachers. This enabled the LETS to be compared to an already validated tool. Peer evaluations were analysed using Generalisability theory; trainee and trainer results were analysed using Classical Test theory.

Results

Following item amendment due to the results of the Cognitive Interviewing process the Delphi process resulted in 19 items allocated to the DOTS and 18 to the LETS. Initial analysis of peer evaluations suggested that the difference in trainer's ability to teach accounted for 44% of the variance in scores between different trainers with 34% of the scores accounted for by the reviewer's natural leniency or harshness. Overall the DOTS showed reasonable reliability with a G co-efficient of 0.44 for one rater; three raters were required for a G co-efficient of 0.7, the generally accepted degree of reliability required for a formative test.

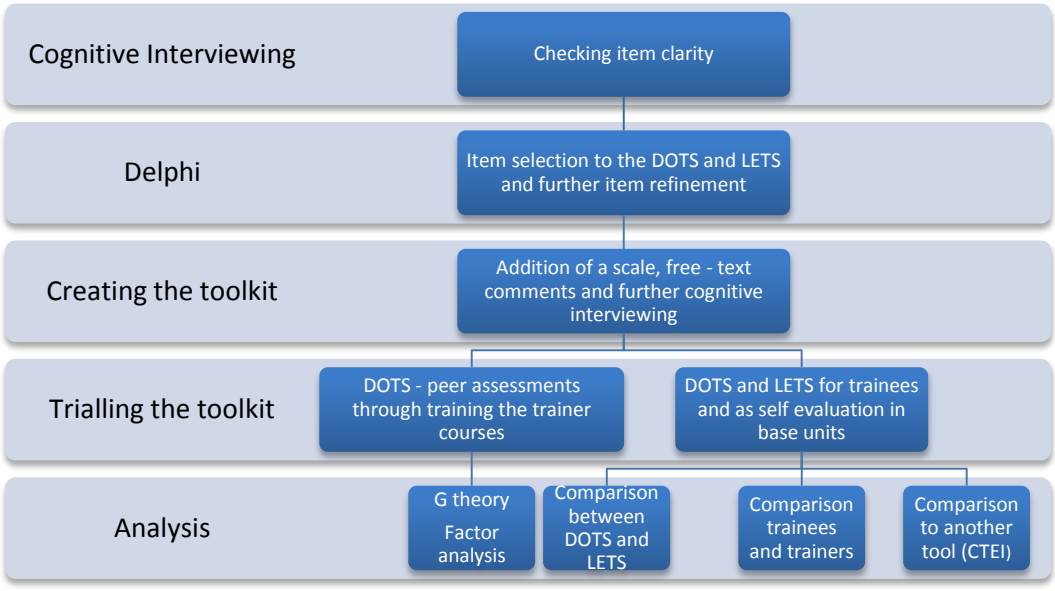
Considering trainee and trainer evaluations the test-retest reliability for the DOTS was 0.75 with a median time length of seven days. Comparing trainee and trainer self-evaluations it was found that trainees tend to evaluate trainers more highly than

trainers evaluated themselves, there was a moderate correlation between trainer and trainee scores of 0.52. The LETS and DOTS showed a high correlation of 0.81 and the LETS correlated highly with the CTEI (r= 0.86).

Discussion

The results of the study are discussed with reference to the American Educational Research Association and American Psychological Association (1999) standards on validity which includes considering the following as sources of evidence for validity; content, response process, internal structure, relationship to other variables and consequences. Content validity was largely contributed to by Wells' (2010) list of attributes but is further developed using the Delphi process. It also continues to map closely to Wells' model of endoscopy teaching as well as established theories of teaching and learning. Response process is partly considered through cognitive interviewing. Reasonable internal structure is demonstrated through the reliability studies. Within local units the DOTS demonstrates reasonable reliability for the DOTS in inter-rater reliability and test- retest reliability. The fact that the LETS correlates highly with the CTEI is strong evidence for the relationship to other variables measuring the same construct. This study did not investigate the consequences of the evaluation tool. This process is depicted in Figure 0-1

Figure 0-1 Flowchart depicting project



Acknowledgments

I wholeheartedly thank my supervisors Professor Roger Barton, Dr Sally Corbett, Dr Chris Wells and Dr Mark Welfare. Their guidance, patience and support throughout this project has been unending and gratefully received.

Thanks also go to Dr Jim Crossley, Sheffield University, who offered statistical advice in the design of the generalisability study.

I am grateful to Northumbria Healthcare Trust for providing the funding for my post and the support from the Research and Development department.

I also thank Professor John Spencer and Dr Richard Thomson for their valuable insights at my university internal assessments.

Finally I would like to give special thanks to all those that gave up their time to participate in this study.

Contents

Table of Figures

List of Tables

Glossary and Abbreviations

Terms used in this dissertation

This dissertation refers to the evaluation of the teaching of gastrointestinal endoscopy. In this field, as with any technical specialty, there is a specific set of terms. To complicate this, several terms can be used interchangeably, I have therefore described what I mean by the words I have used below in order to aim for clarity.

Endoscopist: an individual who is able to use an endoscope to examine the gastrointestinal tract. These individuals come from a variety of backgrounds including physicians, surgeons, nurses and GPs.

Endoscopy: this refers to any gastrointestinal endoscopic procedure. These include upper gastrointestinal endoscopy, colonoscopy, flexible sigmoidoscopy and ERCP. Where appropriate I will refer to the specific procedures as outlined below.

Upper GI Endoscopy: an endoscopic examination of the upper gastrointestinal tract, from oesophagus to the second part of the duodenum. This is also known as an *oesophagogastroduodenoscopy, OGD*, *gastroscopy* or simply an *endoscopy.* This procedure can be either diagnostic or therapeutic. A *diagnostic upper GI endoscopy* may include biopsies (taking a sample of tissue for further analysis) but if any other procedure is required then the endoscopy will become a *therapeutic upper GI endoscopy*. If not specified *upper GI endoscopy* implies a diagnostic procedure.

Colonoscopy: This means an endoscopic examination of the entire colon (large bowel) from caecum to rectum. It can also be referred to as a *lower GI (or gastrointestinal) endoscopy.*

Flexible Sigmoidoscopy: This means an endoscopic examination of the distal part of the colon from splenic flexure to rectum. This procedure is similar to a colonoscopy, but as it examines less of the colon it is less technically demanding. It can also be referred to as a *lower GI (or gastrointestinal) endoscopy.*

ERCP: This is an abbreviation of *endoscopic retrograde cholangiopancreatogram* and refers to an endoscopy that examines the bile ducts using X ray assistance.

The endoscopy community: This term refers to the body of individuals directly involved in delivering gastrointestinal endoscopy service. It includes endoscopists (from all backgrounds), endoscopy nurses (both qualified and health care assistants), reception staff, endoscopy unit managers and trainees in endoscopy.

Trainee: A *trainee* is an individual who is learning to perform endoscopy. They can be from a variety of backgrounds but typically are either specialist registrars in gastroenterology or surgery or nurses.

Trainer: I aim throughout this thesis to refer to anyone who teaches the practice of endoscopy as a trainer; they tend to be Consultant surgeons, gastroenterologists or nurse endoscopists but can also include GPs or other physician specialities.

Expert trainer: in this thesis an expert trainer refers to someone who does not just teach within his or her own local endoscopy unit but also on one of the JAG approved teaching courses (either those designed for trainees or training the trainer courses)

Nurse endoscopist: refers to those that are registered nurses who directly perform endoscopy themselves; as opposed to endoscopy nurses who assist rather than perform the procedure themselves.

Teach & train: These words are similar, as indicated by the dictionary definitions cited below and will be used interchangeably in this dissertation. AskOxford.com defines the verb *to teach* as "1. impart knowledge to or instruct in how to do something, especially in a school or as part of a recognized programme. 2. give instruction in (a subject or skill). 3. cause to learn by example or experience. 4. advocate as a practice or principle." (compact Oxford English dictionary entry – AskOxford.com). It defines *to train* as "1. teach (a person or animal) a particular skill or type of behaviour through regular practice and instruction. 2. be taught in such a way. 3. make or become physically fit through a course of exercise and diet. 4. (train on) point (something) at. 5. make (a plant) grow in a particular direction or into a required shape" (compact Oxford English dictionary entry – AskOxford.com). The practice of endoscopy requires both physical skill and knowledge. Although training implies the transfer of physical skills to a learner and teaching implies the transfer of both knowledge and skills, I do not feel that these distinctions are absolute, for continuity I have opted to use the term train throughout.

Attribute: This research aims to identify the attributes of a high quality trainer of endoscopy. AskOxford.com defines an *attribute* as "1. a characteristic or inherent quality or feature. 2. an object that represents a person, status, or office."(compact Oxford English dictionary entry – AskOxford.com). It is the former definition that I will use to define an attribute. An attribute is a specific quality that is important in that individual's ability to teach endoscopy.

Item: refers to a single question or statement to which a response is expected on a questionnaire or feedback tool. In the case of this study this refers to different attributes of an endoscopy trainer. For clarity the statements describing an endoscopy trainer will be referred to as attributes when discussed not in context of the evaluation tool and items once the tool has been constructed.

Evaluation: this thesis aims to create a tool to evaluate endoscopy trainers. AskOxford.com defines evaluation as 'the making of a judgement about the amount, number, or value of something; assessment'. Evaluation requires those completing the tool to make a judgement with regards to the endoscopy trainers skills. The aim of this project is to attempt to ensure these judgments are reflected appropriately in the results of the tool.

Abbreviations used in this dissertation

DGH	District general hospital
DOPS	Direct observation of procedural skills
DOTS	Direct observation of teaching skills
ERCP	Endoscopic retrograde cholangiopancreatogram
FOB	Faecal occult blood
GI	Gastrointestinal
GP	General Practitioner
GRS	Global rating scale
JAG	The Joint Advisory Group on Gastrointestinal Endoscopy
JETS	JAG endoscopy training system
JRB	Professor Barton (MD supervisor)
LETS	Longterm evaluation of teaching skills
MRW	Dr Welfare (MD supervisor)
N/A	Not applicable
NHS	National Health Service
OGD	Oesophagogastroduodenoscopy
PDP	Personal development plan
SSC	Dr Corbett (MD supervisor)
TTT	Training the trainers
UK	United Kingdom

Chapter 1. Training in Endoscopy

During this chapter I discuss how endoscopy training historically and currently occurs within the UK. I then go onto discuss methods which have been made to attempt to improve training and consider further ways in which improvement could be made. During this discussion I consider the effect of feedback on teachers and consider the various methods which could be sought to measure teaching effectiveness. I conclude this chapter by presenting my aim for this project.

1.1 Training in endoscopy; a historical and current perspective

Digestive endoscopy, a means of visualising the lumen of both the upper and lower gastrointestinal tract, has become a widely utilised tool not only in the identification of gastrointestinal disease but also in its monitoring and treatment. As far back as the 1800s physicians experimented with rigid scopes to visualise the stomach, albeit using professional sword swallowers as their subjects (Sircus 2003). However with the advent of the use of fibre-optics in the 1960s endoscopy moved into more mainstream use and what started as a sideline has become a now huge and complex business (Cotton 2008).

In terms of training the first endoscopists were 'self-taught pioneers'(Cohen 2008) but as techniques progressed and there was a demand for more endoscopists this moved towards a traditional apprenticeship model with experienced endoscopists teaching those who were keen to learn. Originally apprenticeships were largely unstructured but gradually a need to structure this process was acknowledged by the British Society of Gastroenterology and other organisations and hence in 1994 the Joint Advisory Group on Gastrointestinal Endoscopy (or JAG) was created. JAG's (JAG 2011) mission statement is to 'ensure the quality and safety of patient care by defining and maintaining the standards by which endoscopy is practised within the UK'. Today endoscopists come from a wide variety of backgrounds including nursing, medical, surgical and General Practice and all the governing bodies of these groups are represented on the JAG committee. JAG's initial aim was to improve standards for training in endoscopy but its remit is now more wide-reaching and is involved in setting

quality standards and accrediting training units, individual trainees and independent practitioners who wish to perform bowel cancer screening. JAG also quality assured the development of initiated and governed specialised training centres across the UK for residential training courses for trainees and trainers which were initially funded by the Department of Health and are now self-funded.

Over the last decade a major impetus to examine the skills of the UK's endoscopists has come in the form of the National Bowel Cancer Screening Programme (BSCP 2011). This was implemented in 2006 and reached nationwide coverage in 2010. It offers an opportunity for all people aged 60 to 74 to be screened for bowel cancer on a two yearly basis. It involves screening for blood in the stool in asymptomatic individuals. If this is positive the individual is offered a colonoscopy to investigate the cause further. As these colonoscopies are to be performed on otherwise healthy individuals, to be deemed acceptable practice as a screening test they must meet the modified Wilson criteria for screening which state that the test should be valid, reproducible and patient safety must be guaranteed (Longmore, Wilkinson et al. 2001). Bowles (Bowles, Leicester et al. 2004) performed a prospective four month multi-centre UK trial of over 9000 colonoscopies and found that there were large deficiencies in standards including concerns over sedation and consent. In terms of technical aspects complete procedures, defined as reaching the caecum, only occurred in 76.9% and this was reduced to 56.9% if completion was considered to be reaching the terminal ileum. This was well below the completion rate standard of 90% set for trainees. Reasons for incomplete procedure included scope looping and patient discomfort; Bowles (ibid) argued that both of these often reflect poor technique. Bowles also reviewed the training that colonoscopists had received; only 17% had been supervised for their first 100 procedures; below the recommended guideline at that time and only 39.3% had attended a formal colonoscopy course. As Bowles wrote 'the potential of colonoscopy can only be realised if the procedure is completed safely with good visualisation of the mucosa' it is only then it can be considered a 'good' screening test.

The above work highlighted clear deficiencies in endoscopy practice in the UK, and a need to provide better endoscopy services was acknowledged. This was partly in order to provide a bowel cancer screening programme but also to provide all patients with a better service. A variety of measures were put in place in response to these deficiencies with the main focus of improving training for future endoscopy

practitioners and ensuring that all practitioners were appropriately accredited. There were renewed efforts on reinvigorating training courses at three national and seven regional training centres. These training courses each focused on different endoscopic procedures including basic and advanced upper GI endoscopy, colonoscopy and ERCP, and were two to three days in length. A study investigating the effect of endoscopy training courses (Thomas-Gibson, Bassett et al. 2007) found that a five day intensive training course consisting of lectures, simulator and 'hands on' patient teaching sessions improved colonoscopic skills. Trainees attending the course were assessed on their knowledge and skill both on patients and simulators at the beginning of the course, and immediately following the course. Thomas-Gibson (ibid) found that trainees improved in all domains from the beginning to the end of the course. Although JAG-approved training courses are shorter in length than that studied it could be extrapolated that they are likely to have had a similar effect. In a survey of gastroenterology trainees these short courses were found to be valued by trainees (Wells, Inglis et al. 2009).

Along with training courses there have also been advances in technology to assist with training. JAG developed an electronic portfolio; the JAG Endoscopy Training System (JETS) (Mehta, Dowler et al. 2011) which trainees can use to log details of their procedures, develop personal development plans alongside their trainer and receive formative feedback. The JETS system was also used to manage formal assessment of competency and an accreditation process. Electromagnetic scope imagers have been developed which allow both the trainee and trainer to visualise what is happening to the scope and enable the identification of 'loops' within the scope, a particular challenge in endoscopy (Balfour 2001); in use is the ScopeGuide produced by Olympus.

A further technological advance in the teaching of endoscopy was the introduction of simulators to teach; these utilised either animal ex vivo tissue, mechanical colon models or electronic simulators. Simulation has become widespread throughout medicine in recognition that it provides a safe learning environment in which mistakes can occur without harm to patients. Much of this has come from studying the aviation industry who use high fidelity simulators in order to prepare pilots for practice(Bradley, 2006). There are now a wide-variety of models in endoscopy (Cohen 2008) and a myriad of literature supporting their individual validities. Simulators can give trainees further opportunity to improve hand-eye co-ordination and manipulative skills either

prior to or alongside real-life colonoscopy experience. The arguments for the use of simulators were that this would prevent patients undergoing lengthier and potentially more uncomfortable procedures. However, even with the use of simulation devices, feedback must be given alongside the simulator in order for the trainee to improve(Mahmood and Darzi 2004). Although Mahmood (ibid) demonstrated that this can be in the form of computerised feedback, Kruglikova (Kruglikova, Grantcharov et al. 2010) demonstrated that greater improvement was made if this feedback was given by an expert (i.e. a trainer). Additionally simulators have only concentrated on the technical skills of endoscopy. Much has been learnt from the airline industry with regards to the importance of non-technical skills and the use of simulation to explore group dynamics within simulation in order to improve safety within the 'real environment'(Bradley 2006). This has led to developments such as safety checklists in theatre involving the whole team (Flin 2008) This has not yet been developed within endoscopy and an ability to teach both the technical and non-technical skills using simulation may help develop training further. The other disadvantage is that whilst simulators were used to a limited extent on courses for trainees and trainers few units own a simulator. Given this limited access further pursuit of this area is unlikely to improve training throughout all units and therefore has not been pursued further in this thesis.

Despite the above advances the majority of training still occurs within base hospitals and 'hands-on supervised one on one instruction is the mainstay of endoscopy' (Cohen 2008). Several studies have surveyed endoscopy trainees and found that the training within these base hospitals was variable (Bisschops, Wilmer et al. 2002, Wells, Inglis et al. 2009) and, although in a recent re-audit training standards appeared to have improved, 83% of trainees felt it could still be improved further (Haycock, Patel et al. 2010).

In the study looking at the effects of an intensive short training course (Thomas-Gibson, Bassett et al. 2007) the trainees were reassessed at six months following the course and it was found that although there had been no decline in skills there had also been no further improvement. Thomas-Gibson hypothesised that this may have been because the trainees had had very little subsequent colonoscopy experience but also because the training received in their own hospitals was inconsistent or lacking.

Hospital training has faced many challenges; more recently one of these challenges was the change to working hours. Following the introduction of the European Working Time Directives, time for training has been shown to be reduced (Sim, Wrigley et al. 2004) and, although not formally assessed in endoscopy, this is also likely to be true for endoscopy trainees. A reduction in training time means that ensuring the quality of training that a trainee does receive is of paramount importance.

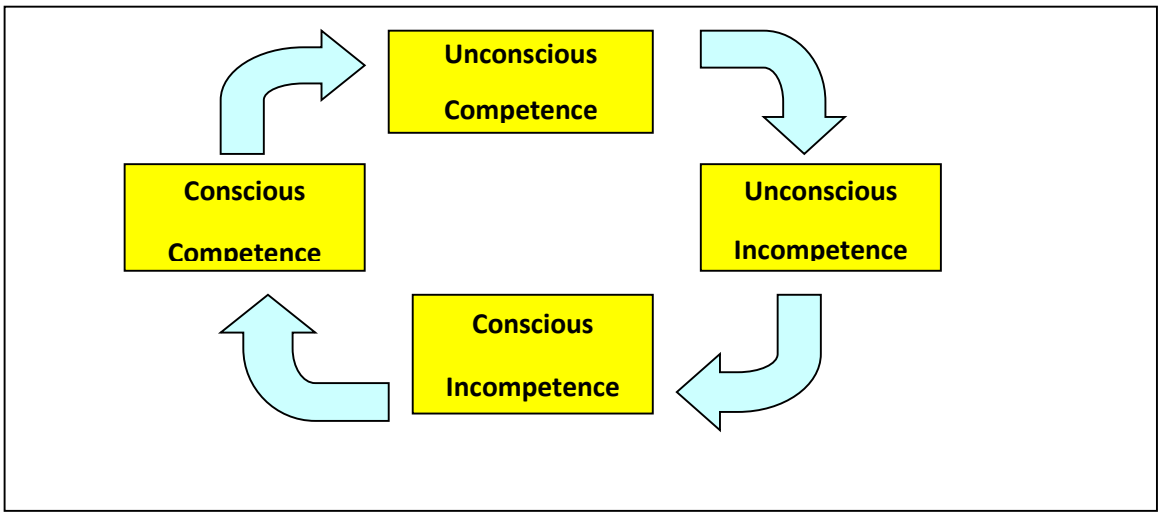
It is apparent from the above discussion that as well as the development of short courses and technologies to assist trainees there also needed to be a focus on improving training within the base hospitals. Some of the changes required were at an organisational level and this need was reflected in the comments made by trainees in the survey of gastroenterology trainees performed by Wells (Wells, Inglis et al. 2009) where trainees made suggestions regarding consultant availability and smaller lists. These organisational factors were assessed by the JAG accreditation system which accredited units for endoscopy and was a mandatory requirement for those units that wish to operate as bowel cancer screening sites. The accreditation process used a system called the GRS (Global Rating Scale) which was a set of standards against which units were expected to audit themselves (JAG). The GRS contained a training domain and set the standards for training within a unit; this included ensuring that there were an adequate number of lists for trainees and that these lists were adjusted to match trainee competence.

1.2 The Endoscopy trainer

JAG therefore had already begun to try and address some of the organisational factors that could improve base hospital training. A further way to improve training within base units was to concentrate on the endoscopy trainer themselves; for as Cohen (2008) states 'the most important ingredient to effective teaching is the teaching skill of the endoscopy instructor' and ultimately training within endoscopy remained a one-to-one process which was seen as the ideal (Teague, Soehendra et al. 2002). One of the challenges for the endoscopy trainer is the difficulty of teaching a complex motor skill; one at which they are now adept at. In the terms of Peyton's (1998) 'learning cycle' the trainer is 'unconsciously competent' (Figure 1-1). This means that they can perform many elements of endoscopy without the need to think through the process, much like driving a car once one has been driving for several years. A new trainee however is unconsciously incompetent (they do not even know the many things they do not know)

but quickly become consciously incompetent. As Teague et al (2002) describe the challenge for the trainer is that they must become again consciously competent in order to breakdown these complex skills in order to train the trainee so that the trainee themselves can become consciously competent.

Figure 1-1 The Learning cycle(Peyton 1998)



As previously mentioned trainers' ability to teach has been found to be variable; this was acknowledged by JAG and in order to aid theses skills acquisition the development 'Training the Trainer' (TTT) courses (JAG 2012) were developed. These were two day courses taught by 'expert' trainers (those with a special interest in training). Day one of the course involved group discussion and practice teaching others in the group using models. The content largely focused on adult learning theory, considering different skills teaching techniques and objective setting. On day two of the course each course participant had the opportunity to teach on a single real case with feedback from the rest of the group. I was unable to find any published data regarding 'Training the trainer' courses, however in the survey performed by Wells (Wells, Inglis et al. 2009) one trainee commented,

> *"All trainers should have a teacher/training plus train the trainer qualification(s) – it is very obvious who has/has not"*

This anecdotally suggests that the courses did create noticeable differences in abilities of different trainers. There is also evidence in other areas of medical education that such faculty development courses do make a difference. A review of 'resident-as-teacher' courses found an improvement in residents' self-rating of their ability to teach

following the course and a small improvement in learner evaluations (Wamsley, Julian et al. 2004). A systematic review of faculty development initiatives (Steinert, Mann et al. 2006) found that there appeared to be positive changes in teachers knowledge, attitudes and skills, though again these were mostly self-perceived changes. The review did find that faculty development programs that used methods with a practical skills-based focus, like the TTT course, tended to be valued by participants. Features that were found to be successful in faculty development programs included experiential learning, providing feedback, involving peers and the use of multi-instructional methods, all of which were utilised in the TTT course. The review found there appeared to be more value in extended programs rather than 'one-off' sessions and this was one disadvantage of the TTT course, however as participants tend to come from all over the country it would be difficult to arrange an on-going program. One possible solution to this would be to arrange local meetings or programs to develop trainer skills further, although this would be difficult to fit into already busy consultant timetables. Additionally the benefits seen from extended programs seemed to largely focus on enabling participants to build networks with other faculty (Steinert, Mann et al. 2006) rather than improving teaching skills per se.

There is some evidence that changes made following faculty development are maintained over time and two studies ((Mahler and Benor 1984, Skeff, Stratos et al. 1986) cited in (Steinert, Mann et al. 2006)) found that changes were maintained up to a year after the intervention; however the review does note that this is an area for further research. So whilst faculty development programs, generally, seem to have some impact it is not possible to say how long this impact lasts. Given this evidence it is likely that the TTT course, as a one-off intervention, has a positive impact on trainer skills. However, even if the changes are maintained there appears to be no impetus for subsequent improvement.

1.3 Methods to provide trainers with feedback

An alternative way of supporting continual improvement of teaching skills of the individual trainer would be to offer feedback to trainers regarding their performance on a regular basis. In order to feedback to a trainer how they are doing in terms of training skills there needs to be a mechanism to measure how good they are as teachers or some way of describing their strengths and weaknesses. This process is normally referred to as evaluation and although this evaluation can occur at any level,

for instance the organisation, curriculum, or program could all be evaluated, I am interested in methods that evaluate the individual trainer.

Clearly the ideal evaluation would be reliable, valid, acceptable by all and inexpensive (Morrison 2003). A variety of methods have been used to evaluate teachers (Snell, Tallett et al. 2000). Berk (Berk 2005) describes evaluation as measuring teaching effectiveness and describes twelve different ways in which this could be done, these are listed in Figure 1-2, and are discussed in turn below.

Figure 1-2. Methods of measuring teaching effectiveness as suggested by Berk (2005)

1. Student ratings
2. Peer ratings
3. Self-evaluations
4. Videos
5. Student interviews
6. Alumni ratings
7. Employer ratings
8. Administrator ratings
9. Teacher scholarship
10. Teaching awards
11. Learning outcome measures
12. Teaching portfolios

1.3.1 Student Ratings

Student views and perceptions of their teachers are most commonly collected using a standardised teaching rating form, which remains the most prevalent method of evaluation (Snell, Tallett et al. 2000, Beckman, Ghosh et al. 2004). Learners experience the teaching first hand and therefore are able to make valid comments about their satisfaction with the teaching or its perceived effectiveness (Snell, Tallett et al. 2000). Student ratings are normally easy to collect and are generally acceptable to teachers; they are normally also inexpensive (Beckman, Lee et al. 2004). Learners experience the teaching first hand and therefore are able to make valid comments about their satisfaction with the teaching or its perceived effectiveness (Snell, Tallett et al. 2000). Student ratings have been collected for a long time both within higher education and medical education, particularly of classroom based teaching. Irby and Rakestraw (1981) were one of the first to consider ratings of teachers within the setting of clinical teaching. They were concerned that because within clinical teaching there existed

greater diversity in the teaching delivered due to the mix of patients that are encountered, the variety of teaching methods and settings used that teaching evaluation would be less reliable compared to classroom teaching. They found however that similar reliability existed between evaluations of clinical teachers using a nine-item tool as had been found in student evaluations of lecturers.

1.3.1.1 Do student ratings make a difference?

Although there are many studies describing different evaluation tools there is limited evidence to the effectiveness of such tools in changing teaching behaviour. In a study which examined the effect of giving teachers feedback gained from student interviews in addition to rating scales (Tiberius, Sackin et al. 1989), the intervention group was compared to a group of teachers who only gained feedback from rating scales and a group of teachers who received no feedback at all. The outcome measure used to measure teaching effectiveness was scores on the rating form completed by subsequent groups of learners. The group which received feedback from student interviews and student ratings showed improvement in their ratings from subsequent groups of students but with respect to the group that received the results of student ratings alone it was found that these teachers did not receive improved ratings. Although it could be argued that there appears to be no benefit in teachers receiving feedback in the form of student ratings it is important to compare this group to the control group who received no feedback at all; this group's ratings actually fell over the period of the study. The authors suggest that this was because the study was performed at a time of flux within the educational system and that it was not unexpected for teachers to feel disenchanted and disengage with the teaching process and therefore in comparison to the group that received no feedback the group which received student ratings seemed to have had some effect.

There are other studies however that appear to show that student ratings do have a definite effect on teaching; however on which teachers this effect is greatest varies from study to study. Litzelman (Litzelman, Stratos et al. 1998) performed a study looking at the effect of student ratings on clinical teachers over one month using a tool based on seven educational categories. The intervention group of teachers was given a baseline summary of their previous ratings, incorporating several years' worth of data,

prior to the intervention. They then received student ratings at the midpoint and then at the end of the rotation. Alongside the summaries of these ratings the teachers in the intervention group also received individual 'teaching effectiveness guidelines' these were individualised reports that emphasised those educational categories in which they were scoring below their peers' average. They also received information about a teaching effectiveness service which could be utilised to aid with advice about teaching techniques; however none in the intervention group utilised this service. The control group had the same summaries of scores produced but these were not made available to them. On analysis of the data it was found that the intervention group with high baseline scores (i.e. were already rated as good teachers) had higher subsequent ratings that the control group with similar baseline scores. In contrast to this however, the teachers who had low initial baseline scores appeared to get worse as their ratings decreased. This study therefore seems to suggest that those who are already seen by students as good teachers further improve with feedback; however those that are seen as more poorly-performing teachers get worse with feedback. This is in contrast to what one might expect given the statistical concept of 'regression to the mean' in which those that receive very good scores initially are likely to score lower at follow up and similarly those with poor scores are likely to improve. Therefore on re-testing those above the mean tend to do slightly worse on retesting and those below the mean scores improve (Streiner and Norman 2008) The authors suggest several possible reasons for this finding; it may be due to the fact poorly performing teachers do not have the necessary skills to effect change or that receiving low scores may make them doubt their capabilities as teachers. One participant suggested that the reason for the further decline in those that scored lower may have been because they became discouraged as they were already trying their best.

Some studies however seem to find that in fact it is the more poorly performing teachers that respond best to feedback from rating forms. In a study of radiology teachers (Cohan, Dunnick et al. 1995), incorporating both their clinical teaching and small-group discussion-style teaching, those that initially scored lowest were those that subsequently improved the most. The study used an evaluation form containing four items which were marked on a ten-point Likert scale. It was completed by residents who had both 'conference' and clinical teaching and the results were fed back to teachers at their individual annual review meeting with the chairperson at the end of the year. This process was repeated the following year and the ratings compared.

Between the two years the mean score for faculty improved in all of the areas examined. The scores for the ten lowest and the ten highest scoring faculty in year one were then examined further. This revealed that the best faculty in year one remained constant in year two. The lowest scoring faculty made a statistically significant improvement in year two, however despite this improvement these faculty remained below the mean score for all faculty. Another study which looked at resident evaluations of surgical trainers (Maker, Lewis et al. 2006) found that the lowest scoring faculty also improved the most. One reason that ratings for the best faculty remained constant in the study of Cohan and colleagues (1995) might be that there was very little room for improvement as the mean score for the ten best scoring faculty were already greater than or equal to nine out of ten on every item making it difficult for the scale to detect further improvement (a ceiling effect). These results are in contrast to those found by Litzelman et al (1998) in that the lower scoring faculty improved and the negative effect of feedback was not seen in these studies. One explanation for this difference may be that the feedback in this study was discussed in person with the teacher whereas teachers in the Litzelman study (ibid) chose not to avail themselves of the opportunity for expert coaching.

Another explanation for improvement of poor performance was proposed (Schum and Yindra 1996) in a study of paediatricians rated by students; those teachers that had received mean ratings for overall teaching effectiveness (one item on the tool) below the departmental mean at baseline showed the greatest improvement by the end of the treatment period. The tool examined ten different domains associated with teaching and found that all faculty who received feedback showed a statistically significant increase in ratings averaged across all ten domains compared to the control group. On examining the domains separately there was a statistically significant improvement in the feedback group in four domains; these were knowledge, demonstrates skill, provides feedback to the trainee and sets reasonable expectations. The authors suggest that the reason improvements were seen in these domains was because it is possible to learn these skills and therefore improve compared to domains which may be more difficult to develop such as rapport. This is a similar argument to the one made by Litzelman et al (1998) that improvement can only be made if the teacher has the methods or skills to develop or change the way they display some attributes within their teaching such as demonstrating their own knowledge more obviously.

All of these studies appear to suggest that giving feedback to faculty in the form of student ratings alone does make a difference to subsequent ratings and may lead to improvements, however there appear to be some inconsistencies as to which teachers benefit from such feedback. Although all these studies relied on student ratings to be fed back to teachers in paper format, there were differences in how this was performed. In some summaries the teacher's results were compared either to the mean score of their peers (Litzelman, Stratos et al. 1998, Maker, Lewis et al. 2006) or were accompanied by comments or advice from those compiling the reports (Schum and Yindra 1996, Litzelman, Stratos et al. 1998) or alongside their own self assessment scores (Tiberius, Sackin et al. 1989, Litzelman, Stratos et al. 1998). In one study (Cohan, Dunnick et al. 1995) these reports were even discussed with the teacher by a senior. These different factors may have influenced the impact of the scores on the teacher and particularly affecting those on whom it had greatest impact. Brinko (1995) also argues that the use of such a facilitator augments feedback to teachers and it has been shown that being given feedback from a person that is trusted and respected can encourage change in areas needed (Menachery, Knight et al. 2006)

Enabling trainees to feedback on their endoscopy trainers was already included in the GRS standards and clearly was felt to be important to JAG; in order for this to occur JAG had created an online trainee evaluation tool which could be found on the trainee's JETS portfolio. This was developed by JAG using a consensus group technique(Fink and Kosecoff 1984) with the group consisting of expert trainers (personal communication with JRB). This tool will be discussed further in the next chapter.

1.3.2 Peer ratings

One option to overcome the perceived disadvantages of student ratings is to collect peer ratings. Peer ratings are seen as being able to overcome some of the perceived disadvantages of student ratings (Speer and Elnicki 1999, Schultz and Latif 2006). In addition to this there is also a hope that they will assist the teacher to develop their own reflective processes about their teaching. Despite this there is limited evidence surrounding the use of peer review in clinical teaching and there is no objective evidence that it improves teaching. There is however a study of the use of peer observation of teaching within a paediatric department which suggests there is subjective improvement in teaching skills(Sullivan, Buckle et al. 2012). Twenty faculty teachers had a teaching session observed by the same observer. The teaching sessions

although clinically orientated were not all necessarily based in a clinical environment. Following the observation the teacher received feedback from the observer and was later asked to email a 'sound-bite' containing their reflection on the process. From analysis of these sound-bites the authors perceived that those who had been observed found it useful and relevant; they strongly valued receiving feedback and felt that it gave them insight and promoted them to reflect on their teaching practices. Teachers also described tangible changes that they had made to their teaching as a result of the observation.

In a survey study and focus group of General Practitioners involved in teaching undergraduates from one university(Adshead, White et al. 2006), regarding their views on peer observation of teaching, there was a general consensus that it would provide a method of addressing problems in their teaching (72 %). The perceived potential benefits largely mirrored those found in the study above, including prompting more reflection on their own teaching and encouragement to try out new teaching methods.

Peer evaluation can be performed either using a rating tool (Siddiqui, Dwyer et al. 2007) or without (Sullivan, Buckle et al. 2012). Beckman (Beckman, Lee et al. 2003) created an evaluation tool to be used by peers in the observation of physicians on teaching ward rounds. The tool showed high internal consistency and good inter-rater reliability. Thirteen of the items on this tool were then used to compare peer and resident (learner) ratings of physicians' teachers(Beckman, Lee et al. 2004). It was found that residents rated physicians higher than the peers on all items, although the difference was only statistically significant in six of the thirteen items. It was also found that peer raters were more reliable than student's ratings i.e. the scores given were more consistent among peers. This supports the argument that peer ratings may be preferable to student ratings as they are statistically more 'sound'.

There are disadvantages to peer observations though. In the survey of General Practitioners (Adshead, White et al. 2006) although the majority felt that peer observation of teaching could be advantageous, over half of those GPs surveyed were not ready to commit to the university's proposed program of peer observation. The reasons cited for this were time pressures, although there was no correlation with volume of workload measures and this viewpoint, or that they felt under scrutiny. The reason for this feeling may have been that the purpose of the evaluation proposed by the University was not only faculty development and teacher support, but also quality

assurance which may have suggested a more summative aspect to the program. This is in contrast with Sullivan et al (2012) who felt their very positive response from faculty was because the formative nature of the observation was stressed.

Peer observation can be time consuming: along with the actual observation it is suggested that the teacher and observer meet both before and after the observation(Siddiqui, Dwyer et al. 2007). Beckman et al (2003) spent nearly a hundred hours in peer-observer time in order to gain three peer evaluations on ten physicians.

There is a tool for peer observation available on the JAG website (JAG 2011); this is used by some units on an ad-hoc basis and is used following TTT courses and training courses for bowel cancer screening courses (personal communication) however it does not appear to be in routine established use nationally.

1.3.3 Self-evaluations

I have already mentioned that some of the studies that utilised student ratings also included as one of their comparative measures teacher self-assessment (Tiberius, Sackin et al. 1989, Litzelman, Stratos et al. 1998). Self-assessment is used widely in medical practice in order to identify learning needs or as part of continuing professional development (Windish, Knight et al. 2004). A systematic review (Davis, Mazmanian et al. 2006) of the accuracy of self-assessment in medicine in comparison to observed measures of competence including others' ratings, found that the conclusions drawn from the 20 studies identified varied considerably. Thirteen of the studies showed little, none or an inverse relationship between self-assessment and other measures whilst seven studies found that there was a positive association between self-assessment and other external measures of competence.

I was able to identify only two studies in the field of medical education that focused on comparing learner ratings with self-ratings. Windish, Knight et al (2004) surveyed physician-teachers regarding their teaching skills and behaviours with each skill marked on a five- point Likert scale; they then asked the same physician-teachers to identify people whom they had taught during the past year. For each teacher two of these learners were then sent the same questionnaire to complete. A comparison of scores revealed that teachers rated themselves lower on all areas assessed; however these higher learner ratings may result from the fact that teachers were able to identify those learners from whom they received evaluations.

A study comparing surgical residents teaching evaluations to the self-perceptions of the surgical attendings who acted as their 'educators' has also been performed (Claridge, Calland et al. 2003). A twenty-item rating tool marked on a five-point Likert scale was created and completed by all surgical residents about all surgical attendings with whom they had had contact. A very similar tool was then distributed to all attendings. Sixty one percent of attendings had given themselves scores that differed significantly from their residents; the authors have not made comment on the direction of the difference but on reviewing the data detailed in the paper this appeared to be a mixture of attendings rating themselves both higher and lower than the residents. The authors also calculated a mean overall score for each attending and a departmental mean. Interestingly those attendings that received mean scores significantly below the departmental mean rated themselves more highly. This concept that those who perform poorly are not aware of this in their own self-assessment has been replicated in other studies (Davis, Mazmanian et al. 2006). Additionally only 78% of attendings completed the self-evaluation exercise. When comparing the residents scores of those attendings who did the self-evaluation and those who did not, those attendings who had not completed it did significantly worse on 17 of the 20 items suggesting that those who perform poorly are less likely to self-evaluate.

The studies described above seem to suggest that perhaps teachers are not very good at completing accurate self-evaluations and that self-evaluation is therefore not a useful exercise. An alternative argument could be that in fact it is the students who are not producing accurate reports. Schwarz and Oyserman (2001) state that what we are doing when we ask a person to comment on someone else's behaviour is gaining a proxy report and that this proxy report tends to be based more on a general impression of the other person rather than an accurate report of their actual behaviours. Due to this phenomenon it has been found that if a tool has a short reference period the degree of convergence between a self and proxy report is low as the self-report is more context dependent however when the time period is increased the reports tend to converge more as both the self and proxy both use dispositional information. Scharwz and Oyserman argue that for short term report of behaviours a self-report is more likely to be an accurate report of the actions that occurred however this is in reference to reporting behaviours such as amount of alcohol consumed in the last week where the self-report is clearly the subject of the tool. I however feel that these distinctions are more blurred in the realm of teaching, although the teacher remains my subject of

interest the student is not just an observer but intimately involved in the process and therefore their different observations may not be more or less accurate, rather influenced by their different perspectives. These differing perspectives and subsequent difference in ratings can be a powerful tool for change (Brinko 1993). Stalmeijer et al. (Stalmeijer, Dolmans et al. 2010) argue that self-assessment enables the teacher to reflect and discrepancies between scores can lead to valuable insights (Berk 2005). This may be useful in those teachers in the above study that had not recognised themselves as below-par teachers.

There is no direct evidence that the use of self-evaluation alone improves teaching. Stalmeijer (Stalmeijer, Dolmans et al. 2010) performed a study examining the perspectives of clinical teachers on self-assessment and, although they felt it helped stimulate reflection, when used alone, the teachers felt that it was of limited use. However, they did feel that discrepancies between self-assessment and the students' evaluations were powerful triggers for reflection and action and agreed that item-level discrepancies were informative. Additionally two of the studies (Tiberius, Sackin et al. 1989, Litzelman, Stratos et al. 1998) that suggested student ratings were effective also included self-assessment and it is not possible to tease out whether self-assessment contributed to the effectiveness.

There is limited literature regarding self-assessment particularly in regard to clinical teaching but there appear to be some good arguments to support its use. Additionally it can be sought in a low cost manner often on the same or very similar tool to that used to collect student ratings. Endoscopy trainers have not had any means by which to complete self-assessment.

1.3.4 Videos

Teachers can be evaluated using video recording of a teaching episode. Any sort of teaching session can be videoed and then reviewed at a later date (Beckman and Frankel 1994). Although video has been around for a long time and is often used as an endpoint for assessing improvement in teaching in faculty development initiatives, there is limited evidence as to whether it improves teaching effectiveness. One study performed in 1972 (Perlberg, Peri et al. 1972) argued that the use of video can bring about change in teaching. Video was used as part of a 'microteaching' exercise, where a teacher is videoed teaching in a simulated classroom, the video was then watched immediately and analysed by the teacher along with a mentor from whom they

received feedback about a specific teaching skill which had been pre-agreed as the focus of the session. The teacher then has the opportunity to make changes to the teaching session and delivered it again. Sixteen study participants took part and underwent the above process once a week for five weeks. The very first video and last video recorded for each participant were rated by a panel of judges on a rating scale and there were found to be changes in teaching style towards a more questioning style in line with the aim of the study. From this study it is not possible to attribute the reason for change in teaching behaviour to the use of video but may instead have been due to the opportunity to receive feedback and re-practise skills in a 'safe' environment. A further disadvantage of this study was that it concentrated on the delivery of lecture skills rather than in the delivery of a clinical session and was performed in the simulated environment of the classroom laboratory.

A study which looks more specifically at the use of video for clinical teaching was performed by Barber(Barber 1992). Six hospital consultants recorded a ten minute teaching session of their choice. The video was then watched by the teacher along with his peers and a tutor who had experience of using video. The video was discussed using Pendleton's rules of feedback (Pendleton 1984 cited in Barber 1992) and the peers then completed two rating scales that assessed the teaching session. This process was then repeated with another video session. An external assessor also watched both videos for each consultant but was blinded to their order and completed the same rating scales as the peers. The consultants who had been videoed completed a questionnaire which looked at their perceptions of their own teaching and the use of video to improve teaching prior to watching the video; they then completed the same questionnaire at the end of the process This was a very small sample group and no statistical analysis was performed on the results but there appeared to be no definite change in either the teacher's perceptions of their teaching or the peer or external assessors rating of their teaching. Despite this it is worth noting that following the session all the teachers felt that using video to improve teaching skills was a useful exercise (although all but one also felt this way at the beginning of the study).

Despite limited evidence for video studies there is clear support for the use of video(Berk 2005), possibly because the object of the focus of the video 'is able to see and hear the data upon which feedback is based'(Beckman and Frankel 1994). Video works best when it uses well-sited equipment in an appropriately sized room and often

it is best when the equipment is permanently sited within a room (Macdougall and O'Halloran 2001). In endoscopy it may be theoretically possible to video teaching sessions as it is based within one room and many of the regional training centres already have the capacity to video within their endoscopy rooms, however this would clearly be expensive to deliver within every unit and beyond many units reach. In addition to this, endoscopy rooms can already be very tight for space in terms of the amount of equipment and staff within the room, therefore it would be challenging in many units to ensure that cameras were appropriately sited in order to ensure that they have good views.

Patients would also require consent to be taken to be videoed. A study of the process of consent for video within the palliative care setting (Hargreaves and Peppiatt 2001) found that whilst nearly all patients did not regret giving permission; 10% felt that the process had been inadequately explained and 6% felt that they had not really understood what they had been asked. Nineteen percent also felt either that they definitely or possibly had not been given sufficient time to consider whether they minded being videoed. These findings are in light of the fact that patients had been sent a letter about the possibility of being videoed a week previously and then discussed it with one of the nurses before signing a consent form, yet a significant proportion of patients still did not believe they had been given enough time or really understood the process. Taking patient consent would equally be a concern within endoscopy as patients are undergoing an intimate examination and are often lightly sedated during the procedure so not fully in control. It would therefore be important to ensure patients were properly informed and consented; this would be pragmatically difficult to do as part of routine practice as patients already need to be appropriately consented for their procedure and the presence of a trainee and therefore the time required to also consent a patient for the use of video would likely be so time consuming that it would be difficult to video on a regular basis.

Additionally it would require the trainer to have the time to watch the episode which may be difficult to arrange especially if a peer were to be present. Due to the limited research regarding its effectiveness and the practical difficulties I do not feel this is a viable option for providing feedback to the base unit trainer within the constraints of this project.

1.3.5 Student interviews

Rather than asking students to complete rating forms their views can be sought in other ways such as interviewing them; these can either be individual interviews or group interviews of students. Tiberius (Tiberius, Sackin et al. 1989) performed a study looking at the effect of group interviews on clinical teachers. These interviews were conducted with 'house teams' that consisted of a medical student, one or two interns and a resident about the senior doctors that had taught them on their current ward over the preceding two months. All interviews were conducted by a group leader and centred on three main questions

- What were the teacher's strengths?
- What were the teacher's weaknesses?
- Can you give some examples

The group discussion was then coded into themes and summarised by the group leader within these themes. The teacher was also interviewed using the same three questions and the discussion summarised using the same codes as the student discussion. The results of these two summaries were then fed back to the teacher using a column format under the code headings, in the first column the teacher's self-perceptions were documented and in the second columns the student perceptions. The authors note that this summary could form part of a discussion between the teacher and the group leader although in this study no discussion took place. As well as completing these group interviews the learners also completed a ratings form. This process was repeated over four successive cycles of different learners and it was found that ratings improved (compared to a group of teachers who only received learner ratings for whom no successive improvement in ratings was seen). This study therefore shows that student interviews appeared to improve teaching.

This would be difficult to replicate within the field of endoscopy training for several reasons. One would need to allocate a group leader potentially to every hospital. This group leader would need a considerable amount of training, not only in interview techniques but also in the coding of interviews and producing the summaries. Time would need to be allocated for the group leader to conduct the interviews but also time to analyse the interviews and produce the summaries. As the group leader would require training it would be likely that only a few would take on this role therefore

would be required to conduct this process for several endoscopy trainers. Both the learners and teachers need interviewing which may be difficult to arrange and as often a trainer may only be training one trainee at any one time individual, rather than group, interviews would need to be arranged. Although this study does show that student interviews can bring about improvements in teaching it would not be a practical method to currently give feedback to all endoscopy trainers.

1.3.6 Alumini Ratings

As well as current students, course alumni could complete rating forms. One concern about alumini completing feedback forms is that they may not be able to remember the course in detail. However it has been shown that there is a high correlation between current students ratings and alumni for up to four-years post graduation (Overall and Marsh 1980 cited in Berk 1995). But if these ratings are so similar to current students ratings one could argue then what is the point in collecting both sets of data? Berk (1995) suggests that graduates may be able to contribute a different perspective in terms of preparedness for work and can also comment on topics such as timing of certain subjects and curriculum content. As I am keen to provide feedback to individual trainers these latter subjects are of less interest as they can be beyond the individual endoscopy trainers' control as the curriculum is not set by the individual trainer. Learning endoscopy takes several years throughout which a trainee is likely to have had several different trainers and as trainees tend to rotate around different hospitals this training is likely to have taken place in several different units. Once practising independently ex-trainees may have different insights to those that they had whilst training but it may be difficult to attribute it directly to a single trainer as it may be the input of several trainers or due to the set-up of a particular unit. Furthermore feedback would clearly be delayed and therefore whilst useful has little immediacy and may not occur regularly.

1.3.7 Employer or administrator ratings

Berk (1995) and Snell (2000) both argue that it is important to look beyond the expected or 'standard' sources of evaluation evidence. This includes using administrators or employers to rate teachers; Berk argues that they can offer an alternative viewpoint. He argues that employers of previous students can offer an opinion on students 'readiness for work' however as endoscopy tends to be learnt over several years in different hospitals with different trainers it would be difficult to use

this method to evaluate a single teacher. Additionally many of the skills that an employer may value as worthwhile such as time-management, communication, team-working are not just exhibited within endoscopy training as these are generic professional skills. Administrator ratings are often based on documentation completed by the trainer rather than direct observation (ibid) for all elements of their job not just teaching hours and may evaluate aspects over which the trainer has little personal control. For example, the amount of time an endoscopy trainer spends training is often dictated by the endoscopy unit as JAG recommend that training should occur on specially shortened lists to allow time for training. Evaluating this would provide little useful information about the actual trainer's performance and the feedback could not be actively used by a trainer to improve the training they deliver.

1.3.8 Teaching scholarship and teaching awards

Berk also suggests that teaching scholarship and teaching awards can be used to measure teaching effectiveness. In terms of teaching scholarship Berk is referring to measuring teaching effectiveness by the number of publications and presentations a teacher makes. He argues that this is a surrogate for teaching expertise; however I feel this only indicates an interest or expertise in educational research rather than in teaching itself. This is also likely to be unhelpful in measuring teaching effectiveness in endoscopy as many trainers are not educationalists within their own right and therefore very few indeed are likely to have such a research history. This is also true of teaching awards; these are not routinely awarded in endoscopy and therefore could not be used as a current measure. In addition to this who receives teaching awards is dependent on the selection process of that award and therefore is highly dependent on the validity of that selection process.

1.3.9 Learning outcome measures

If one agrees with the maxim that ' the ultimate criterion of good teaching is learning (pg439)' (Iwaszkiewicz, DaRosa et al. 2008) then a logical argument would be to use learning outcomes as a measure of teaching effectiveness. Using learning outcome measures means judging the effectiveness of teachers based on how their students perform. A study of surgical students' performance in examinations compared with quality of teaching has been performed(Blue, III et al. 1999). The students had recently completed two four week surgical rotations for which they had been assigned a faculty member for each rotation who acted as their preceptor throughout that rotation. The

preceptor was responsible for the student's educational experience throughout the month and was rated by the student at the end of the rotation on a three item three point scale. The mean result of this rating scale was then compared to the student's marks in a written paper and the data gathering and data interpretation components of an OSCE. The study found, having corrected for the student's previous academic performance that mean score of the rating scale was significantly associated with student performance on the written examination. The authors also identified the top 20% of teachers with the highest rating scores and labelled these 'best' and the 20% of teachers with the lowest ratings (labelled 'worst'). If they looked at student performance compared to these categories of teachers they found that students who had had one of the 'worst' teachers performed significantly less well on the data gathering station of the OSCE. The authors use these results to argue that this demonstrates that high quality teaching does make a difference to students but could also be used as an argument that as teaching quality and learning outcomes appear to be linked these could also be used as a measure of teaching quality. However improvement is not seen in all areas of academic performance; the authors do not suggest why this might have been for instance it would be unfair to measure teaching effectiveness against data interpretation in this study group as no correlation with teaching performance was found. Additionally the preceptor was responsible for the student's educational experience but did not necessarily deliver all of this experience themselves; it could be hypothesised that these students had received excellent teaching elsewhere throughout their rotation which made them consider their supervisor more favourably but is not actually a measure of the teachers skill. Although previous academic performance was taken into account based on previous examination scores it does not appear that these were solely on the subject of surgery. It may be that those students who wish to pursue a career in surgery were more positive about the rotation and therefore applied themselves more in examination. Additionally those who enjoyed the rotation as a whole may have been more positive in evaluating their supervisor and also more motivated to learn for the examination process. Clearly it is difficult to separate out what factors may effect a student's performance; student performance is not necessarily directed linked to teaching effectiveness and includes many other intrinsic factors within the student. This is also difficult within endoscopy as it has been increasingly realised that competence in endoscopy is partly related to the number of procedures performed but also that

every trainee is different in terms of how quickly they become competent. This process can be aided by a trainer but can never just be the whole responsibility of the trainer. In addition to this trainees do not sit an exam or a test every year and learning endoscopy can be a slow process and therefore a measurable milestone might not be reached in every rotation or with every trainer. Trainees are encouraged to undergo regular formative assessment but this assessment is performed by their own trainers and therefore it would be inappropriate to use this as a measure of teaching effectiveness as it may affect how a trainer scored a trainee.

1.3.10 Teaching Portfolio

Another possible method of measuring teaching effectiveness is to use a teaching portfolio. A teaching portfolio is supposed to be an amalgamation of several sources of evidence and may also include a reflective component (Berk 2005). If endoscopy trainers were asked to produce a teaching portfolio that teaching portfolio would need to contain a variety of evidence that demonstrated their effectiveness as teachers. This evidence could document the hours that they have spent training and the details of those they have taught, however this sounds very summative in nature and would do little to inform a trainer how they might improve their teaching practice. A portfolio would therefore need to contain some of the above sources of evidence that would help measure teaching effectiveness that would inform the trainer in what ways they might improve. A key area in the construction of a portfolio for the GMC's process of revalidation (GMC 2012) is that not only must the portfolio contain evidence in the named areas but also that this is a reflective process and that the doctor should reflect on each of these areas. If a trainer were to keep a teaching portfolio, such reflection may help identify areas for improvement.

1.3.11 OSTE

One method of measuring teaching effectiveness not suggested by Berk which has been advanced in recent years is the Objective Structured Teaching Exercise (or encounter) (OSTE). The OSTE was first described in 1992 and follows a similar format to the OSCE (objective structured clinical examination) used to assess students in that they both follow a standardised format (Trowbridge, Snydman et al. 2011). The OSTE essentially consists of a simulated teaching scenario using a standardised student, who is trained to react in the same manner to every teacher, and a trained observer or peer. Following the teaching episode there is often the opportunity for immediate feedback

from both the student and the peer (Boillat, Bethune et al. 2012). The OSTE has therefore been used as part of a faculty development program either as part of the program itself (Boillat, Bethune et al. 2012) or to assess a program's effectiveness (Stone, Mazor et al. 2002) or as a method to evaluate teaching effectiveness either formatively or summatively (Trowbridge, Snydman et al. 2011).

In a systematic review of the OSTE (Trowbridge, Snydman et al. 2011) it appears to show reasonable reliability with good inter-rater agreement and validity in terms of the fact that most participants seem to find it realistic. In terms of whether the OSTE improves teacher skills most studies have looked at teacher perceptions of this area which tend to be positive but there is less objective evidence as to whether there is objective improvement.

An OSTE can therefore provide an effective means to evaluate teachers which is normally well received by teachers themselves. Clearly an OSTE has large resource implications but also in the field of endoscopy training it would be difficult to standardise the scenario if it were to contain real procedures as each case is different and therefore it would be impossible for the trainee to behave in the same way in every situation. An OSTE could be developed in which trainers taught on simulators but then this may not be as realistic or acceptable to trainers. An OSTE is an artificial method by which to evaluate teaching in that it does not evaluate teaching within the actual workplace and therefore has the potential to lack ecological validity. Ecological validity refers to the concept, normally in reference to experimentation or testing, that the process should match real life as much as possible(Cohen, Manion et al. 2007) It is important to try and match the characteristics and factors of a given situation as closely as possible in order to extrapolate the results. As an OSTE is done under a false setting then this may reduce its ecological validity particularly if one decided that each case must be the same and a simulator was used.

Another problem with OSTEs identified by several authors (Morrison, Boker et al. 2002, Trowbridge, Snydman et al. 2011, Boillat, Bethune et al. 2012) is that a rating form is required; often this rating form has not been validated and therefore this brings the validity of the OSTE under scrutiny.

1.3.12 Summary

One of the main features emphasised by many authors (Irby 1983, Berk 2005, Siddiqui, Dwyer et al. 2007)is that one single method to evaluate teachers is inadequate as each method has its own limitations and will not give a rounded view of that teacher. Several methods should therefore be used; this process is called triangulation (Bye, Connolly et al. 2007). Triangulation enables any inadequacies in one method to be compensated for by the other methods in order to gain a fuller picture. Irby (1983) discusses how at one medical school in order to overcome this problem that no single method of evaluation is perfect several methods were integrated to create what he believed to be a robust method to evaluate teachers. This included student, peer and self-evaluation, although peer evaluation did not necessarily refer to peer observation of the teaching itself rather to peer review of documentation surrounding the teaching process.

In order to give effective feedback to endoscopy trainers therefore, we should consider utilising more than one evaluation method. Above I have discussed many different methods of evaluation, some of these I believe, given the reasons described in their individual sections, have limited use in trying to give feedback to endoscopy trainers, these include alumni, employer or administrator ratings; teaching scholarship or awards; and learning outcome measures. Whilst teaching portfolios might be useful consideration needs to be given to what might populate such a portfolio. One of the major advantages of portfolios is to try to stimulate reflection(Mathers, Challis et al. 1999)but they also need to contain concrete evidence of teaching (Lamki and Marchand 2006). This could just be a list of teaching hours but can include evaluations which could then be used to stimulate reflection; therefore whilst a teaching portfolio for endoscopy trainers could be useful in further developing teaching it is currently necessary to develop other methods which could eventually populate such a portfolio. Trainee interviews could provide plentiful information about a trainer's performance however these would be difficult to arrange and the interviewer would require training. A different interviewer would need to be provided for each trust and this would require lots of interviewers to be trained and an interview structure devised. This may be a useful measure to further try and explore those trainers that appear to be struggling but would not be practical to receive regular feedback to all trainers. This therefore leaves trainee ratings, peer ratings, self-assessment, use of videos and the OSTE. All of these methods do have one similar concept in that they all commonly use a rating

instrument in order to record an opinion and to structure feedback. Clearly both videos and OSTE are more expensive both financially and in terms of organisation to set up, and it would be difficult to provide such services nationwide to improve the base hospital trainer in every trust and therefore are less preferable than the other methods. One of the components of triangulation is that there should be overlap between the different methods (Jahangiri, Mucciolo et al. 2008) therefore one argument could be that it would be useful to create an evaluation tool that could be used by trainers, trainees as a self-assessment and peers; this would allow a rounded view of trainers to be developed and optimise the feedback that trainers receive. Potentially the same tool could then be used in future to analyse videos or as a rating scale on an OSTE.

1.4 Aim

The aim of this study therefore is to create an evaluation tool that can be used to give feedback to an endoscopy trainer by a trainee, a trainer as a self-assessment and a peer.

Chapter 2. Psychometrics of Evaluations and Review of Other Tools

In this chapter I consider other evaluation tools in existence and consider how the validity of an evaluation tool can be judged. I initially performed a literature search to identify all endoscopy evaluation trainer tools, review articles for clinical evaluation tools and surgical evaluation tools. I then consider what evidence can be sought for validity and then consider the studies I found in light of this.

2.1 Evaluation tools in the literature

At the end of the last chapter I introduced my aim to create an evaluation tool that could be completed by peers, trainees or by the trainer as a self-assessment exercise. Streiner and Norman (2008) write that the most common error made by clinical researchers is to write new scales and reject old scales too easily misjudging the complexity that designing a new scale requires; 'therefore a useful first step is to be aware of any existing scales that suit the purpose. The next step is to understand and apply the criteria for judging the usefulness of a particular scale (pg.5)' (Streiner and Norman 2008). In this chapter I will consider tools that already exist that may be suitable for evaluating endoscopy trainers.

In order to identify appropriate papers I performed three literature searches. Initially I wanted to identify whether any endoscopy trainer evaluation tools already existed. I already knew of one tool that was in current use on the JETS website. In order to identify any other tools I performed a literature search of MEDLINE, EMBASE and ERIC from 1946 to July 2012 using the following search terms

- Endoscopy, digestive system OR Endoscopy, gastrointestinal OR Endoscopist
  AND
- Teacher OR Teaching OR Faculty OR Trainer
  AND
- Evaluation OR Feedback Or Effectiveness

This identified fifteen citations, the abstracts for these were all reviewed but none identified tools to evaluate endoscopy trainers. I therefore decided to consider clinical teacher evaluation tools. The phrase 'clinical teaching' is used frequently throughout

the literature but is not often defined. It tends to be used to refer to teaching that occurs 'on the job' rather than in lecture or classroom based settings with patients present (Ramani and Leinster 2008). This means that the teaching environment can be unpredictable and the teacher has to take into account not only the student and their learning but also the patient and the time pressures of the working environment (Spencer 2003). Endoscopy training fits under this umbrella of clinical teaching albeit a rather specific subsection and therefore I felt that studies that looked at this wider setting may also be relevant to endoscopy training. I therefore performed a further literature search of MEDLINE, EMBASE and ERIC from 1946 to July 2012 using the following search terms but limited the search to review articles.

- Medicine OR medical OR clinical OR bedside
  AND
- Teacher OR Teaching OR Faculty OR Trainer
  AND
- Evaluation OR Feedback Or Effectiveness

This identified three review articles written in English (Beckman, Ghosh et al. 2004, Beckman, Cook et al. 2005, Fluit, Bolhuis et al. 2010) that reviewed instruments used to evaluate clinical teachers.

I then opted to only examine surgical tools more closely. The reason I chose to look at surgical training is that as mentioned above endoscopy teaching is a specific form of clinical teaching in that it is teaching a complex skill. In surgery the trainer is also teaching a complex procedural skill as well as all the other skills associated with being a doctor or health professional. I hypothesised that this teaching of a complex procedure would change the needs of a trainee and therefore trainer attributes outwith that of normal clinical teaching and therefore it would be most likely that a surgical tool would be more relevant to endoscopy teaching than other evaluation tools. I searched MEDLINE, EMBASE and ERIC from 1946 to July 2012 using the following search terms

- Surgery OR surgeon OR surgical

AND

- Teacher OR teaching OR faculty OR trainer

AND

- Evaluation OR feedback OR effectiveness

This identified 144 citations. Titles and abstracts were reviewed for suitability and for those that appeared relevant the whole paper was acquired and read in detail. Reference lists of the above papers were also used to look for other relevant studies. I was interested in those studies that evaluated the individual teacher rather than the program or course as a whole. I was also only interested in those studies that utilised some sort of rating tool and described the development of the tool. I therefore excluded any tools that evaluated the programme rather than the teacher or were pre-clinical in nature. This identified twelve papers that described nine different tools which concentrated on the individual trainer (Downing, English et al. 1983, Tortolani, Risucci et al. 1991, Risucci, Lutsky et al. 1992, Cohen, MacRae et al. 1996, Hauge, Wanzek et al. 2001, Cox and Swanson 2002, Claridge, Calland et al. 2003, Maker, Curtis et al. 2004, Sarker, Vincent et al. 2005, Iwaszkiewicz, DaRosa et al. 2008) although in one study the questions that addressed the individual teacher were part of a longer tool that also looked at other aspects of the surgical training program (Cohen, MacRae et al. 1996).

Once I had identified potentially suitable evaluation tools it was then necessary to consider on what basis I would judge their quality as an evaluation tool. Streiner and Norman (2008) also state that it is necessary 'to understand and apply the criteria for judging the usefulness of a particular scale (pg.5)' this is referred to as the psychometrics of evaluation. I will consider what such psychometric data can be considered prior to considering the tools.

2.2 The psychometrics of evaluation

Psychometrics traditionally refers to the 'science of psychological assessment' (Rust and Golombok 2009) but is now commonly used in education and clinical contexts when subjective measures are relied upon. Psychometrics enables one to consider how a rating tool should be constructed, what properties it should be tested against and provides criteria by which it can be judged; this section discusses those properties.

2.2.1 Validity

Validity refers to the evidence given to support or refute the meaning or interpretation given to the results of an instrument (Downing 2003). In other words providing validity evidence is about providing the evidence to demonstrate that an instrument measures what it purports to measure i.e. the construct under investigation. In this case this

would be providing evidence that the tools can identify good and bad teaching skills. Proving or disproving validity is similar to hypothesis testing in the basic sciences in that the proposed interpretation should be stated and then evidence gathered to support or refute this interpretation until either this interpretation is felt to be plausible or has been rejected (Cook and Beckman 2006).

There are two key concepts which must be recognised when discussing validity; one is that validity refers to the meaning given to the results of an instrument rather than the instrument itself (Cook and Beckman 2006). The reason for this distinction is that validity is situation specific; for instance, just because a tool evaluating teaching is deemed to have good validity evidence in the classroom it does not necessarily mean that this evidence would support its use in a ward setting. This is important when considering the surgical and clinical evaluation tools as clearly the evidence presented for validity is not necessarily evidence for their validity when used to evaluate endoscopy trainers if it has not previously been used for this population. The second key concept of validity is that an assessment or evaluation can never be said to be valid or invalid but is a continuum in which evidence is provided to either support or refute the proposed interpretation of scores(Downing 2003).

There are several different ways in which validity evidence can be sought and categorised. One method is to divide validity evidence into three categories; content, construct and criterion-related (DeVon, Block et al. 2007). More recently it has been suggested that all validity can be viewed under the heading of construct validity (Downing 2003, Cook and Beckman 2006). This viewpoint stems from the fact that many tools are attempting to measure constructs that cannot easily be measured such as teaching effectiveness or professionalism. These constructs are 'intangible collections of abstract concepts and principles which are explained by educational or psychological theory (pg. 831)'(Downing 2003). Due to the fact that the construct cannot be quantitatively measured, validity should try to demonstrate that the tool is, in a surrogate way, measuring the intended construct. Validity can therefore be seen as a single concept but evidence from multiple sources can be used. The American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education in their joint publication "Standards for Educational and Psychological Testing" (American Educational Research Association, American Psychological Association et al. 1999) agreed with this unifying concept of

validity and suggested five areas in which evidence to support validity can be found (Downing 2003, Beckman, Cook et al. 2005, Cook and Beckman 2006, Fluit, Bolhuis et al. 2010). These are listed below and are then explored in further detail in relation to the evaluation of teaching.

1. Content
2. Response process
3. Internal Structure
4. Relationship to other variables
5. Consequences

2.2.1.1 Content

Content validity refers to ensuring that the content of the evaluation matches the construct under investigation. This can be relatively easily demonstrated in student assessments where the content of an assessment can be compared to the curriculum. In terms of teaching this is more difficult to demonstrate as clearly there is no curriculum that states what makes a good teacher. Sutkin, Wagner et al (2008) argue that every individual within the field has some personal concept of what they feel makes a good or bad teacher but that there is no over-arching view. They therefore performed a review of the published literature that described attributes relevant to good clinical teaching. They reviewed 68 articles and within these found 480 descriptions of characteristics of good clinical teachers, which they grouped into 49 themes. They then divided these attributes into cognitive and non-cognitive attributes and found that the majority of descriptions given described non-cognitive attributes. This in contrast to what is taught on faculty development courses which tend to focus on the cognitive attributes of technical skills. This leads to a dilemma as to what should be measured on an evaluation tool as it suggests that mirroring what is taught on a faculty course might not accurately represent the many attributes that make a good teacher. It would also not be feasibly possible to utilise all 480 descriptions of characteristics. Sutkin, Wagner et al (2008) conclude that 'superb teaching is certainly a complex phenomenon (pg.458)'; this complexity can lead to difficulty when considering the content validity of a teaching evaluation tool as the concept is not

clearly defined and there is no blueprint of teaching which can be matched to the items.

As a result of this a variety of methods to derive and demonstrate that items cover a sufficient view of 'the good teacher' are used and it is important that a discussion about how these decisions have been made are documented. Methods used include using relevant literature to decide which attributes to include(Clarke 1999, Keely, Oppenheimer et al. 2010), although the discussion above demonstrates that this can be difficult given the myriad of characteristics found by Sutkin, Wagner et al (2008). Other methods are adapting previous tools(Beckman, Lee et al. 2003) and using stakeholders to help devise or refine items. These stakeholders can be 'experts'(Copeland and Hewson 2000), faculty (Dolmans, Wolfhagen et al. 2004, McGrath, Yeung et al. 2005) or the learners (Donner-Banzhoff, Merle et al. 2003). Choice of stakeholders may change the selection of items; for instance faculty and learners have been shown to place importance on different teacher attributes; learners tend to place greater importance on interpersonal skills whereas faculty place greater importance on technical skills such as punctuality and organisation (McLean 2001).

In their review Sutkin, Wagner et al (2008) first organised the descriptions of the characteristics into themes and then organised these themes into three categories; teacher characteristics, physician characteristics and human characteristics. These categories help us to review the attributes but do not necessarily provide us with a theory that helps us understand how these attributes build and interact to form a good clinical teacher. Identification of an underlying theory can be useful in item development as it enables the researcher to understand how the different attributes interact and their respective importance; the resulting tool should then be developed utilising all areas of a relevant theory or model (Streiner and Norman 2008 pg 21).

2.2.1.2 Response process

The response process is examined to ensure that the method in which the tool is administered does not alter or affect the interpretation of the results. This can include using a format that those using the tool are already familiar with, and ensuring the instructions for completion are appropriate. The aim of considering the response process when designing a tool is to try and reduce error associated with test administration as much as possible (Downing 2003). This includes considering the

wording of the items but also how participants will answer those items. This normally includes the use of a scale, which can affect the responses given by participants, for instance the wording of the scale and the number of points on the scale can all affect how the participant answers (Schwarz and Oyserman 2001).

As part of the response process one should also consider how the results are presented, for instance if the results are presented as combined scores, either in domains or as a total score, then it should be ensured that the method by which the scores are combined is appropriate. This category can also include ensuring that those completing the tool are interpreting the questions appropriately and as the authors of the tool intended; Cook and Beckman (2006) suggested this could be checked by asking students to 'think aloud' as they answered questions about teachers.

2.2.1.3 Internal Structure (and reliability)

This category refers to the statistical properties of the items in the instrument(Downing 2003). One method to do this is to calculate the internal consistency of the instrument; items on an instrument which are intended to measure the same construct should correlate more highly.

The reliability of the results is also considered as contributing to the evidence for internal structure. Reliability refers to the reproducibility of results (Beckman, Ghosh et al. 2004). It is the 'degree to which a result reflects all possible measurements of the same construct(pg 802)' (Crossley, Humphris et al. 2002) and is a way of quantifying the amount of error which is seen in any measurement (Streiner and Norman 2008).

When any tool is administered the aim is to measure how much or how little a person exhibits or possesses the construct under investigation; this is referred to as their true score. However the true score is not the score that is actually seen on the test, as there will always be an element of error; therefore the score on the test is termed the observed score. This concept is also represented in Equation 2.2 where X symbolises the observed score, T the true score and E the error (Rust and Golombok 2009).

Equation 2.1. Equation to explain concept of true score where X = observed score, T=true score, E =error

$$X = T + E$$

It is important to note that the true score can never be known for certain, as all error will never be eliminated and therefore the true score can only ever be estimated.

When calculating the reliability of a rating tool we are attempting to numerically state how close the observed score is to the true score; worded differently reliability is telling us how large the error term is. As reliability examines the variability in scores it can therefore be expressed using the equation in Equation 2.2 (Crossley, Davies et al. 2002).

Equation 2.2. General formula for reliability

$$R = \frac{true\ variance}{total\ variance}$$

$$R = \frac{true\ variance}{(true\ variance + error\ variance)}$$

There are different methods by which the reliability co-efficient can be calculated. These include Classical Test theory and Generalisability theory where Classical Test theory examines possible sources of error individually in contrast to Generalisability theory which examines all possible sources of error in the same analysis. Like validity there is no absolute figure that means that reliability has been reached as it can depend on many factors including the purpose for which the instrument is being used in the first place. There is general agreement though that the higher the stakes of an instrument then the more reproducible the researcher would want the results to be. Downing(Downing 2004) suggests that for a high-stakes exam one would want a reliability of 0.90 whereas if the instrument were to be used for more formative purposes a reliability of greater than 0.70 would be considered acceptable.

Examining the reliability of a tool is essential to validity; if a tool produces completely different results every time or for every rater then those results become difficult to interpret and use to inform practice. It is important to note, however, that reliability alone is not enough evidence to support validity (hence the other categories of validity evidence also mentioned in this chapter). Reliability is therefore seen as one component that contributes to validity evidence; albeit a very important component. In a similar way to the concept of validity, reliability refers to a specific test situation and cannot automatically be generalised beyond that setting; due to this it is the reliability of the scores that are under review rather than the tool itself (Streiner and Norman 2008).

2.2.1.4 Relationship to other variables

The purpose of the 'relationship to other variables' category is to examine how the tool compares to other tools or measurable behaviours (Cook and Beckman 2006). For instance; an alternative instrument can be applied to the population under investigation; if the instrument claims to measure a similar construct a high correlation would be expected, for example a similar teacher evaluation tool; this would be termed confirmatory evidence. If the instrument measures a different construct one would not expect them to correlate and this could count as counter-confirmatory evidence(Downing 2003). If looking for confirmatory evidence one could argue if the two instruments correlate too highly then why develop a new instrument but this could be because the new instrument is shorter or quicker to administer or is to be used for a different purpose; for instance one might want to use a tool that contains more detail if it is for formative use than a short highly reliable tool for summative use that provides the teacher with less detailed feedback.

2.2.1.5 Consequences

This is the most controversial category of validity evidence(Cook and Beckman 2006) and looks at the effect that an instrument has on those subjects it is used on. It considers what happens as a result of the scores given and attempts to investigate whether there are any unintended effects. The argument is that no harm should come from an assessment or at the very least more positive than negative should arise as a result of the test (Downing 2003). For instance in Section 1.3.1 I discussed the study performed by Litzelman et al (1998) that looked at the harmful and beneficial effects of giving feedback. This study found that those teachers who had low baseline scores did worse at follow up than controls that also had low baseline scores despite the fact that they had received interim student feedback. This suggests that feedback had a detrimental effect on their teaching. This could be seen as a negative consequence of the evaluation tool and although it does not necessarily mean that the tool should not be used, the organisation might need to consider how the feedback is delivered to teachers to ensure that it does not have a negative effect.

2.2.2 Summary

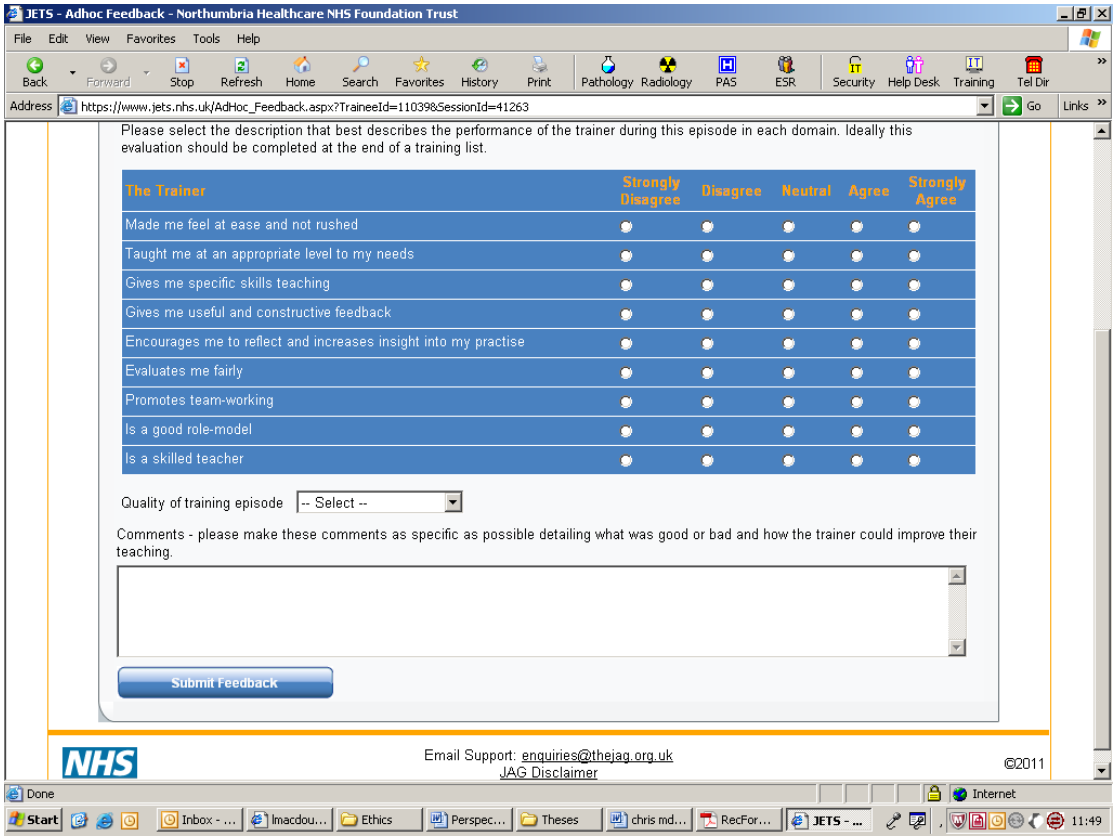
The above discussion reflects the various methods and types of evidence that can be used to contribute to reliability and validity evidence when considering the possible sources of evidence for validity as set out in the "Standards for Educational and Psychological Testing" (American Educational Research Association, American Psychological Association et al. 1999) . The next section will review those tools identified by my literature search with reference to the evidence presented for validity and their suitability for the evaluation of endoscopy trainers.

2.3 Endoscopy evaluation tools

As previously mentioned a tool designed to evaluate endoscopy trainers by trainees already exists and can be found on the JETS website (JAG 2012) and is displayed in

. It was created using a consensus group technique(Fink and Kosecoff 1984) with the group consisting of 'expert' trainers (personal communication JRB).

Figure 2-1 Screen shot of JAG endoscopy tool for trainers to be completed by trainees



Content validity is often judged qualitatively rather than quantitatively (Rust and Golombok 2009); on first glance the items all appear reasonable as characteristics one might expect an endoscopy trainer to possess but these may not represent all the attributes of the endoscopy trainer and it is not possible just by examining the tool to determine whether these items represent the most important attributes. As the tool was created by a group of 'expert' trainers then the resulting tool only reflects their view of endoscopy training. I have already discussed that teachers and learners might emphasise different attributes of a teacher as most desirable (McLean 2001). It is therefore arguably important to include trainees in the development of any tool as other evaluation tools have done (Donner-Banzhoff, Merle et al. 2003) The 'expert' trainers that were involved included those that taught or were involved in the JAG-approved 'Training the Trainer' courses, by incorporating only their views there is a danger of just reproducing that which is already taught without considering other perspectives; creating 'cultural reproduction' (Bourdieu cited in (Moore 2000)).

Apart from an acceptance that 'expert' trainers thought these were the most important or relevant characteristics for evaluating endoscopy trainers there is no further evidence to support Content as a part of evidence for validity. In addition there is no evidence that these forms are reliable and measure what they purport to be measuring; there is no evidence for any of the other sources of validity for this toolkit. This is not to say that this tool is not valid but there is no evidence to support its validity. It would be possible to collect data regarding the reliability of this tool which would provide evidence for its validity however the content would continue to reflect the views of expert trainers only. As previously mentioned there was also a peer tool on the JAG website, again there is no supplementary information about this tool's performance in terms of psychometrics.

There are no other published tools of endoscopy training. As there are limitations with the above tool it is necessary to look outside the field of endoscopy to examine whether there are any other existing tools that could be utilised to give formative feedback to endoscopy trainers and to further consider what evidence can be provided for validity. I therefore felt that it was necessary to consider tools that had been empirically tested for evidence of their psychometric properties. I initially opted to consider tools that evaluated clinical teachers.

2.4 Clinical Teacher Evaluation Instruments

There are already three comprehensive review articles that investigate the validity of instruments used to evaluate clinical teachers (Beckman, Ghosh et al. 2004, Beckman, Cook et al. 2005, Fluit, Bolhuis et al. 2010). All three reviews looked at clinical teaching but the studies reviewed varied in their site of teaching, for instance whether it was in-patient or out-patient or both, the level of the reviewers (students or junior doctors), and the speciality of the teacher, with general medicine the most commonly evaluated speciality (Beckman, Ghosh et al. 2004). All tools used a rating scale with the number of items varying from one to 43 items and all used a Likert scale ranging from four to ten points (Beckman, Ghosh et al. 2004).

In 2004 Beckman et al found that every tool studied did produce some evidence to support validity using the standards set out above; however the type of evidence provided varied from paper to paper. The most common source of validity evidence was Internal Structure; with the most common statistics used being either factor analysis or demonstrating internal consistency by calculating Cronbach's alpha. In terms of content the authors grouped the domains within the tools studied to look for commonalities within the actual content of the tool. They found that 14 different domains of teaching were used; the two most common groupings of items that occurred across the tools were 'clinical teaching' and 'interpersonal'. They cite evidence that students seem to be able to distinguish between these domains (Donnelly and Woolliscroft 1989) and hypothesise that tools that are based on these two domains alone may be adequate.

Beckman et al (2005) then reviewed the same studies (plus one further study) but in order to try and quantify the amount of evidence used for validity they used a rating scale for each area of validity in which

N = no discussion of this area of validity

0 = discussed but no data presented or data failed to support validity

1 = data for this source weakly supports the validity of score interpretation

2 = data for this source strongly supported

They then reviewed each of the possible sources of validity evidence using these criteria against the 22 studies. No paper scored a two for all sources of evidence. As each study was marked out of a possible two for each area this means that each source of validity evidence could have scored a possible 44. Using the scoring system the authors determined that internal structure was the most commonly presented evidence for validity (scoring 32 out of 44) followed by content validity. These were the only two categories where some studies scored two out of two (except for one paper which scored a two in relation to other variables). Consequences and Response Process were the least common sources presented as evidence of validity (in fact each scored only two out of a possible 44).

Each study was rated by two of the reviewers; the degree of agreement between these two reviewers was calculated using the kappa co-efficient. It was found that there was good to excellent agreement for Content, Internal Structure and Relation to Other Variables however there was poor agreement between raters for the categories Response Process and Consequences. The authors hypothesised that this may be because the former were more common and therefore were easier to identify as patterns emerged. It may also be that it is more difficult to display concrete evidence for areas such as Response Process and Consequences and therefore the evidence given is open to greater interpretation creating less inter-rater agreement. This review demonstrates that no studies produced high levels of evidence for every validity category. This may be because such information has not been published and articles regarding these tools focus on one component part. The fact that evidence in all these areas is not available or not published may also represent the fact that collecting evidence for all sources of validity represents a substantial amount of work that is both labour intensive and requires evidence to be collected over a long time period. The areas which were more represented are those that can be performed ad hoc on data collected particularly in terms of the internal structure with often no a priori design requirements. This is in comparison to areas such as Relation to Other Variables which would need to be considered at the start of the development process in order to ensure that the correct data is collected.

Although the scale used in this study is helpful in ascertaining that some sources of validity evidence are underrepresented in the literature there is a danger that one could use such a scale to say that one tool is more valid than another. This is not

necessarily the case as it may depend on the purpose of the tool. Additionally although Beckman et al (2005) score evidence for the content validity of the toolkit based on the description of how the items were derived they do not look at the items themselves. They therefore do not comment on whether the items do capture the construct of clinical teaching or not.

Fluit et al (2010) looked more recently at the validity evidence provided for instruments to evaluate clinical teaching and reviewed 33 different tools. Similar to Beckman they found that internal structure was the most common area for which evidence was provided. Fluit et al (2010) however examined the content of the tools more closely; they reviewed the literature of the characteristics of good teachers and derived attribute domains from the literature. These domains were physician role model, teacher role, supervisor role (assigning work and feedback), supportive person, assessor role and planner organiser. Fluit et al (2010) found that 30 of the 33 instruments contained an assessment of the teacher role, followed by 29 assessing the supporter person, 27 the role model and 26 the feedback element of the supervisor role. These four domains accounted for 79% of all the items on all the tools.

Fluit el al (2010) also argued that an important component of being a clinical teacher is that the teacher acts as a role model to the learner. By acting as a role model they should display the desired attributes of a physician. In order to investigate this they compared the instruments against the Canadian Medical Educational Directives (CanMEDS) which describe the competencies of a physician as medical expert, communicator, collaborator, manager, health advocate, scholar and professional. A third of the items on all the tools could be related to these competencies although more than half of these were related to the medical expert and scholarship.

From these reviews it is possible to see that although the American Psychological and Education Research Associations suggest five areas for which evidence of validity can be sought that in reality the field from which evidence is sought appears to be narrower in most studies. Fluit et al. (2010) demonstrate that the content of many tools does not take into account all the different domains that one might expect of a good clinical teacher. The reasons for this might be varied and include the fact that the tool may not be intending to measure all these areas. This is in keeping with the fact that instruments tend not to be generalisable as institutions have their own culture of teaching and the tool should reflect that culture (Snell, Tallett et al. 2000). The other

reason why these tools may not include all the aspects of being either a doctor or a teacher is that in order for attributes to be included on a toolkit they need to be measurable and some of these areas may be difficult to describe in measurable ways. Therefore in trying to evaluate endoscopy trainers it may be possible that one of these tools that is designed to measure the attributes of a good clinical teacher would also be suitable for the evaluation of an endoscopy trainer but this may mean that important attributes specific to endoscopy are missed and therefore the tool may not measure the appropriate construct. I therefore opted to only examine surgical tools more closely. The reason I chose to look at surgical training is that in surgery the trainer is also teaching a complex procedural skill as well as all the other skills associated with being a doctor or health professional. I hypothesised that this teaching of a complex procedure would change the needs of a trainee and therefore trainer attributes outwith that of normal clinical teaching.

2.5 Surgical evaluation instruments

As discussed in the introduction, nine different surgical tools were identified that were described in 12 different published articles.

2.5.1 Content validity

Given that one of the concerns over the endoscopy tool in terms of its content validity was the process by which the items had been derived, my interest when considering the evidence for content was twofold –I was interested in how the items had been derived and whether the items would be suitable for also evaluating an endoscopy trainer, i.e. are they attributes that I would expect an endoscopy trainer to display in the course of their teaching.

One study made no mention of how the items on the tool had been derived (Cohen, MacRae et al. 1996). In the other studies the details were often limited to a few lines or a short paragraph. Methods included performing a review of the relevant literature (Cox and Swanson 2002, Sarker, Vincent et al. 2005) and/or involving a mixture of residents, medical students, teaching faculty or educators. One of the tools used the same criteria on which their residents were judged (Tortolani, Risucci et al. 1991, Risucci, Lutsky et al. 1992). Two tools involved only the learner in selecting the items (Maker, Curtis et al. 2004, Iwaszkiewicz, DaRosa et al. 2008) and five tools involved both the learner and the teacher in their development (Downing, English et al. 1983,

Hauge, Wanzek et al. 2001, Cox and Swanson 2002, Claridge, Calland et al. 2003, Sarker, Vincent et al. 2005). The five tools that involved the learner and teacher do not fully describe how these two groups' views were utilised to form the tool. Methods mentioned include individual discussions with trainees and trainers (Sarker, Vincent et al. 2005); via 'consultation' with faculty and residents (Hauge, Wanzek et al. 2001); using their 'opinions' (Cox and Swanson 2002) or using an 'ad-hoc committee' (Claridge, Calland et al. 2003) but no further information is given to how these views were then amalgamated and incorporated to create the final tool. One of the two tools that used learner opinions (Maker, Curtis et al. 2004) stated that they asked senior residents to collaboratively decide on nine characteristics that described a surgical role model but does not inform us of how the residents worked in collaboration to make these decisions. The other study that utilises learners' views (Iwaszkiewicz, DaRosa et al. 2008) stated that the medical students and residents were surveyed to identify teaching behaviours associated with being an outstanding teacher within the operating room; this process is described in another paper however I was unable to source this paper including the abstract.

In terms of the actual content of the tool all of the tools used a Likert-type rating scale with between 3 and 5 points. The number of items varied from four (as part of a longer tool) and 26 items. Three of the studies just concentrated on the interaction that occurred in the operating room (Hauge, Wanzek et al. 2001, Sarker, Vincent et al. 2005, Iwaszkiewicz, DaRosa et al. 2008) whereas others also looked at behaviour in the ward and clinic setting either separately (Cox and Swanson 2002, Claridge, Calland et al. 2003) or as combined behaviours over a duration of a rotation (Downing, English et al. 1983, Tortolani, Risucci et al. 1991, Cohen, MacRae et al. 1996, Maker, Curtis et al. 2004)

The items have been summarised in Table 2-1 under what I believe to be appropriate headings, these are headings that I designed in order to group the items and to be able to easily visualise the different items that had been included in the tools; they are not the only interpretation of how the items could be grouped. As can be seen from the table the tools vary in terms of their content. Some of this difference is likely related to the fact that the tools were designed for use in different areas and demonstrates the need to be aware of a tool's purpose in order to make an assessment of the appropriateness of the content, for instance those that are to be used in the operating

room contained a higher level of detail about this area. The most commonly assessed attribute in seven of the nine tools is feedback. Additionally nearly all the tools make reference to the need for good communication between the trainer and trainee. In addition to verbal communication, five of the nine tools also make reference to the trainer's ability to demonstrate a skill appropriately; this is clearly important when teaching a practical skill.

Table 2-1. A table of the attributes evaluated in surgical trainer evaluation tools. The number denotes the paper that that item can be found. 1= Sarker, Vincent et al. 2005; 2= Cohen, MacRae et al 2006; 3= Hauge, Wanzek et al. 2000; 4=Cox, Swanson 2002; 5= Claridge, Calland et al 2003; 6= Maker, Curtis et al 2004a, 7= Iwaszkiewicz, DaRosa et al 2008; 8= Downing,English et al 1983; 9= Tortolani, Risucci et al 1991 and Risucci, Lutsky et al 1992;

Within the operating room - setting the scene	
Explains surgery about to happen	1, 3, 4, 5
Discusses likely patient outcome and possible complications	4, 5
Sets out the aims and responsibilities of the trainee	1, 4, 5
Outlines when trainer will take over	1
Within the operating room - Scaffolding	
Pays attention to surgery	1
Gives trainee reasonable time to complete surgery	1
Allows trainee to continue if makes mistakes	1
Gives trainee space within surgical field	1
Allows trainee appropriate autonomy	1, 6
Permits resident participation in surgery according to ability	4, 5, 6, 8
Awareness and sensitivity to trainee's learning needs	4, 5
Demonstration/ discussion	
Demonstrates task appropriately	1, 3, 4, 5, 6
Demonstrates decision making appropriately	6, 7
Allows learners to feel pathology	7
Other clinical areas	
Performs didactic teaching	6
Attends didactic teaching	6
Makes significant contributions to resident's learning at conferences	8
Teaching rounds	6, 8
Assists resident to find and complete research for publication	6
Co-operates in all aspects of trainee's surgical education	8
Gives resident opportunity to teach	4, 5
Communication	
Ability to challenge thinking and encourage resident to think critically	2, 4, 5
Stimulate critical thinking with use of literature	6
Ability to communicate	2
Answer questions clearly	2, 3, 4, 5, 7
Explains tasks appropriately	1, 3
Sets tone appropriately	3
Encourages residents questions	4, 5
Personal attributes	
Confident in role as teacher and surgeon	7, 8

Role model	8
Reliability	9
Respect for patient	7, 8, 9
Shows good judgement	8
Reaction to pressure	9
Personal appearance	9
Has up-to-date knowledge	4, 5, 8, 9
Remains calm	7
Exhibits fairness	7
Interpersonal attributes	
Creates a climate of mutual respect for all in team	4, 5, 7, 9
Respect for patient	7, 8, 9
Awareness and sensitivity to trainee's learning needs	4, 5
Feedback	
Provide useful feedback	1, 2, 3, 4, 5, 6, 7
Makes a future plan for trainee	1
Attitude towards teaching	
Attitude towards teaching	2, 7
Creates positive learning atmosphere	4, 5, 8

Within the table some of these attributes could have been amalgamated further, for instance one could argue that reaction to pressure and remains calm are really measuring the same attribute, however this may not be what the authors intended. In relation to whether any of these tools could be utilised to evaluate endoscopy trainers, in terms of the items themselves, the vast majority of endoscopy teaching occurs within the endoscopy unit which is not dissimilar to the operating room therefore many of the skills that are included that make reference to other clinical areas of work would not be relevant to the endoscopy trainer. Those tools that look at the operating room appear to only concentrate on technical skills however an endoscopy trainer does not just need to teach the trainee to do the procedure but also interact with the patient and staff. This concept will be investigated further in the next chapter.

Cox et al (2000) also collected qualitative comments asking residents to pass comment on the teaching strengths of each surgeon as well as complete their rating scale. During the analysis stage they coded these comments to identify recurrent themes. Three main themes were apparent; demonstrates technical expertise, allows resident participation and maintains a learning climate of respect. The authors argue that all these themes were also represented in their rating scale and therefore could be argued that matching these open comments to the themes within the items is evidence of the content validity of the items. None of the tools specifically made reference to any theory of teaching or learning when discussing the derivation of the items.

2.5.2 Response process

There is limited evidence that response process has been considered in the development of these surgical evaluation tools. All tools use a Likert scale for their rating scale consisting of either three to five points but no tool makes reference to why these numbers of items were chosen. One of the tools was completed electronically (Iwaszkiewicz, DaRosa et al. 2008) whilst the others appear to have been completed on paper. By the prevalence of such rating scales there appears to be an assumption that these are acceptable to learners and that they know how to complete such tools. In order to help learners score their teachers Sarker et al (2005) did add descriptors to aid their decision making. They also comment that 'all trainees thought the assessment tool was relevant and clear(pg 418)' (Sarker, Vincent et al. 2005) but they do not explain how this was assessed. Hauge et al (2002) used an observer who had been trained to complete all the evaluations; this training is likely to have ensured that the tool was completed as intended.

The timing of when the respondents complete a tool will also affect how they respond; the longer the time period from the actual event then respondents may have forgotten specific behaviours and responses are more likely to rate general disposition (Schwarz and Oyserman 2001); alternatively they may have had time to reflect on the experience and this may alter their responses. The tools that consider a single teaching session, for example a single operating list, (Hauge, Wanzek et al. 2001, Sarker, Vincent et al. 2005) were completed either during an event by an observer or immediately after the event if completed by a trainee. The tools that evaluated teaching over the duration of a rotation were often completed immediately at the end of the rotation (Downing, English et al. 1983, Cohen, MacRae et al. 1996). For some of the tools it was unclear from the description when exactly the evaluations occurred.

The response process requires consideration of how scores are combined and the results fed back to the subjects under scrutiny; it also considers the time period between the evaluation and the feedback given to subjects. Despite this six of the studies did not report how the results were fed back to the teachers (Risucci, Lutsky et al. 1992, Hauge, Wanzek et al. 2001, Cox and Swanson 2002, Claridge, Calland et al. 2003, Sarker, Vincent et al. 2005, Iwaszkiewicz, DaRosa et al. 2008). Four studies described how the results were fed back to the teachers. For all of these tools feedback was anonymous and trainers were only told about their own performance although

head of faculty often also reviewed every faculty member's feedback. Three of the studies fed back a score for every item (either as a mean or all scores) (Downing, English et al. 1983, Tortolani, Risucci et al. 1991); this means that the trainers received the maximum amount of detail that the tool allowed for as they were able to review how they had performed on every item; this may be useful if the purpose of the tool is to try and initiate change . In contrast Cohen et al (1996) whose tool was created for summative use combined the total scores for the four items pertaining to teaching effectiveness and gave teachers a mean total score. Although these two methods largely appear to match their intended use in that you would expect more detail from a tool that was intended for formative use, there is no justification given for these methods of feedback or their acceptability to the teachers involved. As well as the scores themselves the teachers were often given some idea of how they performed in comparison to their colleagues; often this was in a graphical form where their position in relation to their colleagues was demonstrated (Cohen, MacRae et al. 1996, Iwaszkiewicz, DaRosa et al. 2008). In terms of timing the feedback was given either on an annual or biannual basis (Downing, English et al. 1983, Cohen, MacRae et al. 1996, Cox and Swanson 2002, Iwaszkiewicz, DaRosa et al. 2008), normally as a written report but in one study those that received lower scores were required to meet with the vice-chair for education to discuss methods for improvement (Iwaszkiewicz, DaRosa et al. 2008).

2.5.3 Internal structure

Internal structure is the most commonly expressed evidence given to represent validity (Fluit, Bolhuis et al. 2010); eight of the tools presented data for internal structure. Six of the studies considered the internal consistency of the tool. Maker et al (Maker, Lewis et al. 2006) performed part-whole correlations looking at the correlation between each item and the mean total score. Four of the studies (Downing, English et al. 1983, Hauge, Wanzek et al. 2001, Cox and Swanson 2002, Iwaszkiewicz, DaRosa et al. 2008) used Cronbach's alpha to examine the internal consistency of the tool with three of the studies quoting a Cronbach's alpha of greater than 0.85 (Downing, English et al. 1983, Hauge, Wanzek et al. 2001, Cox and Swanson 2002) and the fourth study quoting a Cronbach's alpha of 0.78(Iwaszkiewicz, DaRosa et al. 2008). These studies either present this data without any further interpretation of its meaning or comment that this shows high internal consistency. Alpha is a function of both the degree of correlation between items and the number of items within the tool (Field 2009). The

above mentioned tools contained between ten and 20 items, Hauge et al describe alphas for the individual sections of their tool which vary in item length from four to ten items. This differing number of items affects the interpretation of Cronbach's alpha. One study (Tortolani, Risucci et al. 1991)also performed factor analysis which can give further information about the structure of a tool. The requirements and process of factor analysis are described further in chapter 7 but in this study the data fits the requirements for factor analysis but there is no information about how the factors have been extracted.

Two studies used more than one rater type, a trainee and a trained observer (Hauge, Wanzek et al. 2001, Sarker, Vincent et al. 2005). In these studies the inter-rater reliability was calculated and presented as an inter-observer agreement (86 -97%) or a k co-efficient of 0.77 where the k co-efficient measures the amount of inter-rater agreement but also takes into account that any agreement may be due to chance. As well as the agreement between two different types of raters the amount of agreement between the same type of raters can also be calculated. An intraclass correlation was calculated in one study to examine the degree of agreement between trainees (Cohen, MacRae et al. 1996) and Risucci et al created a trainer-rater matrix to compute mean inter-rater correlations.

2.5.4 Relationship to other variables

No study compared their tool to another evaluation tool that purports to measure the same construct. Claridge et al (2003) asked residents and trainers (as a self-assessment exercise) to complete the same rating tool. They found that 61% of attendings (the teachers) scored themselves significantly differently from the residents (the learners). Although two different variables are being measured and evaluated in this study; the fact that they correlate poorly does not necessarily mean that the tool was invalid but it could suggest that the construct of self-perception of teaching is a different construct to that of learner perception of teaching; Schwarz and Oyserman (2001) discuss that self and others reports of behaviour may differ.

Maker et al (2004) compared the scores of the nine separate items in their tool to one global question which asks the learner about the surgeon as a role model in which the learner could answer that "I don't want to emulate", "OK", or "a Role Model" (Maker, Curtis et al. 2004). They found that each of the other nine more descriptive items on the tool correlated significantly with the Role Model category; they also found that

three of the items were uniquely associated with the role model variable; this comparison suggests that the tool is measuring a similar construct. Similarly Iwaskiewicz et al. (2008) found that there was a correlation between their items that looked at teaching skills shown in the operating room with an overall teaching effectiveness score.

Tortolani et al. (1991) compared the ratings received from residents with other activities of teaching and surgical practice. In terms of teaching practice they compared scores to the number of major procedures performed with residents, the number of Grand Round and Morbidity and Mortality Conferences attended and the number of research articles published in the last three years. They found a significant difference in ratings between those that practiced more of all the above except attendance at Grand Rounds. The authors therefore argue that those who engage in more teaching practice and research are more highly rated by residents.

2.5.5 Consequences

Three studies looked at the consequences that using the tool had upon teachers. Cohen et al (1996) looked at the use of their tool over a nine year period and found that the mean score was stable over that time. The primary aim of their tool was to recognise those teachers that should be acknowledged for promotion; they found that once promoted their teaching effectiveness score declined slightly, which was in contrast to those who were initially found to be poor teachers whose score improved. Maker et al (2004) sent the results of their evaluations to surgeons along with a personalised narrative; they then repeated the evaluations six months later and found that the average score for each of their 9 items had improved, with this improvement reaching statistical significance for three items. Maker et al (2004) found that the lowest scoring faculty initially seemed to benefit most from being evaluated. They then evaluated faculty again one year on from the previous evaluation and 18 months since the first evaluation and found that faculty ratings continued to improve (Maker, Lewis et al. 2006). Downing et al (1983) also found that the lowest scoring surgeons seemed to improve most when given their feedback.

2.6 Discussion

The above shows how different studies present evidence for validity of surgical evaluation tools in different ways. No tool presented evidence for validity in all five

categories. It is important to note that this only includes the published information about these tools; these tools may have been investigated in other ways internally. In the articles published they may have only been aiming to highlight one aspect of the tool for example the focus of one study was a comparison of trainee evaluations with self-assessment (Claridge, Calland et al. 2003). This mirrors what was found in the reviews of studies of clinical teacher evaluations (Beckman, Cook et al. 2005, Fluit, Bolhuis et al. 2010). Evidence for the content of the tool and internal structure are more frequently discussed compared to the other sources. The possible reasons for this are discussed in Section 2.4 but include the fact that it is easier to examine this data without too much consideration of the study design. Another reason that different emphasis is placed on different sources of validity evidence may relate to the different purposes for the tools.

2.6.1 The purpose of evaluation

Evaluation tools are created for different purposes; purposes listed in the literature include improving teaching, to provide encouragement for teachers and to support applications for promotion for teachers (Morrison 2003). There is no level at which a tool can be said to be valid or reliable; rather it is an ongoing spectrum. As can be seen from the different reliability co-efficients proposed by Downing (2003) the level of reliability required depends on the purpose for which the tool is being used. Acknowledging the purpose is also important when it comes to deciding in which categories evidence of validity should be sought (Fluit, Bolhuis et al. 2010).

In the UK audit or evaluation of teaching is deemed good practice by the General Medical Council (GMC 1999) but, aside from this, evaluation of teachers can be performed for a variety of purposes as listed above. Therefore evaluation, like assessment, can be viewed as falling into either summative or formative categories where summative assessment requires students, or in this case teachers, to 'demonstrate the "sum" of their knowledge, skills and/or attitudes'(Rolfe and McPherson 1995) and is normally associated with a pass/ fail or certification decision. In the case of teaching this can be viewed as evaluating teachers for the purpose of awarding promotions, bonuses or awards. Formative assessment, in contrast, is to enable students to assess their current level of understanding and knowledge and promote development. From the point of view of evaluating teaching, formative evaluation is to provide encouragement and improve teaching practices.

Acknowledging the purpose of evaluation is important in the development of a tool as different uses of the results will alter its desired psychometric properties (Downing 2003).

Differences in formative and summative evaluation have been described with reference to assessment of the learners but the principles can be applied to evaluation of teachers. Harlen and James(Harlen and James 1997) emphasised the importance that there is a clear distinction between summative and formative assessment. They argue that as deep learning occurs through building on prior knowledge it is essential for both the teacher and the student to identify where the student currently is in terms of their knowledge and that this is the purpose of formative assessment. They also state that in order to improve the student's deep learning the student themselves should also be aware of where they are in terms of their knowledge and skills. They conclude therefore that formative assessment is essentially feedback 'both to the teacher and the pupil about present understanding and skill development in order to determine the way forward'(Harlen and James 1997).

If formative assessment is used in this diagnostic way then Harlen and James (ibid) warn that it may appear contradictory as students are often changing and may appear to be able to understand or complete a task in one setting but not in another. If this were a summative assessment then this would be frowned upon as the assessment would be seen to be unreliable, however Harlen and James (ibid) argue that the fact the student appears not to be able to translate these skills from one setting to another is actually useful to the teacher as it provides information as to further learning needs of the student and it is this feedback that is vital to formative assessment. Therefore they argue that in formative assessment one should not be overly concerned with the reliability of the test but that validity is essential because one needs to be measuring the correct construct in order to inform further learning.

In terms of teaching evaluation, if used formatively to improve clinical teaching the instrument should provide relevant feedback about the teacher's strengths and weaknesses; if used for promotion or ranking it should be able to distinguish between good and bad teachers in a highly reliable way (Fluit, Bolhuis et al. 2010). For instance a global score for teaching can identify those faculty that students perceive as strong or weak but, without more feedback as to why, it is not possible for a teacher to know what they specifically need to change (Iwaszkiewicz, DaRosa et al. 2008)

I stated at the end of chapter 1 that I intended to create an evaluation tool for endoscopy trainers that could be used formatively. My intention was to try and provide a mechanism by which they could understand and reflect on what their current strengths and weaknesses as teachers are in order to further improve. To this end I need to ensure that the tool contains enough detail in order to adequately describe a trainer's current performance. For high stakes summative evaluation the reproducibility of results is particularly important whereas for my formative evaluation the ability to provide an accurate description of the desired characteristics within a context may be seen as most important as this may allow the teacher greater knowledge about their current teaching and a further awareness of how to improve. That does not mean that my formative tool should have no reproducibility or that a summative tool cannot provide useful feedback to an individual but highlights that different emphasis may be put on these factors.

Schuwirth and van der Vleuten (2006) (Schuwirth and Vleuten 2006) also state that caution must be taken when using psychometrics to make decisions about what constitutes good or bad assessment. They state a similar position to Harlen and James (1997) in that the construct under investigation may not actually be stable and can change from situation to situation, for instance a high degree of knowledge in one area does not necessarily mean that the student will have a high degree of knowledge in another. They argue that using statistical models that measure variance or correlations results in the rejection of large amounts of information which may provide useful information about the student. Schuwirth and van der Vleuten (2006) do not make the distinction between formative and summative assessment but one can see that for formative assessment discarding this information would not be in the student or teacher's interest as it may provide information about student deficiencies that need addressing. Psychometrics, especially in terms of reliability, appears to make those who develop assessments strive to create homogeneity however variance reflects the medical world in terms of difference in cases and teaching settings and assessment should reflect that.

In reference to the purpose of the surgical tools the tool's purpose was not always clearly stated within the literature; many do point out that these evaluation tools can be used to feedback to teachers. Sarker et al (2005) state more specifically that the aim of their tool, which concentrates on the teaching of technical skills in the operating

room, is for formative evaluation. They also state that by developing a tool they hope to create a recognised process by which technical skills should be taught. Maker et al (2004) also suggest that their tool can be used to recognise teaching discrepancies and be a basis for improvement. Some of the tools are designed for summative use, in order to measure who the best teachers are and award promotion; this can either be the primary purpose of the tool (Cohen, MacRae et al. 1996) or Cox et al (2002) suggest their tool can be used optionally for promotion by including it in a teaching portfolio. The tool described by Hauge et al (2001) has a different purpose; the tool as it is described in the literature was not discussed in terms of whether it can measure attributes that are indicative of good teaching but whether actual moments of teaching can be identified in the operating room and whether those moments could be noted reliably for the future purpose of researching or assessing teaching in this environment.

The difference purpose of the tool also affects the acceptability of the method by which the results of the tool are fed back to the teachers evaluated. As discussed in Section 2.5.2 the different surgical tools used different methods to provide feedback to teachers. One just gave a summary score (Cohen et al 1996) whereas other tools (Downing, English et al. 1983, Tortolani, Risucci et al. 1991) fed back a score for every item. If the tool is for summative use a summary score may be acceptable however if it is formative use then the teacher requires more information in order to develop and improve. Acknowledging the purpose of the tool is therefore important in both the design of the tool and its testing evidence of validity in order to ensure it is fit for purpose. Tools can however be used for both purposes; currently within endoscopy training trainees are expected to complete the same tool for both formative and summative assessments. It is used throughout the process of learning a procedure such as OGD in order to guide progress and gain feedback on how they are doing. In this way it is used formatively to guide progress and aid development. As the same tool is also used summatively it guides trainees as to when they might be ready to undergo summative assessment. The trainee has to sit several summative assessments, if these are passed, along with certain other criteria, they are deemed competent to practice independently.

2.6.2 Other aspects of validity

The American standards for validity are now well established and useful in exploring the evidence provided by tools for their validity. Beckman et al (2005) created a scale by which to judge each source of validity. Although the scale was helpful in ascertaining that some areas were better represented than others it does not necessarily mean that a tool that scored lower than another is less valid because as discussed above the interpretation of evidence of validity can be affected by the purpose of the tool, which was not taken into account within the scale. Additionally the two reviewers did not always agree on the score given to each tool, which suggests that opinions will differ as to what counts as evidence and the strength of evidence.

The 'Standards' are also only one way to view validity and although are well established this is not to say that this is the only way that validity or indeed evaluation tools per se can be judged. For instance traditionally one type of validity that was considered was face validity; this refers to the acceptability of a test or evaluation method (Rust and Golombok 2009). This can refer to the method of assessment or evaluation used, for example, that it does not offend or embarrass anyone but also that the test is taken seriously. In terms of teacher evaluation this could include the level to which teachers believe that the feedback from such a tool is meaningful. Centra (1993) describes conditions in which teaching evaluation is likely to improve teaching, these include ensuring that teachers gain new knowledge regarding themselves or their performance from the evaluation, that it should be from a credible source, that they are provided with information about how to change, and that faculty have motivation to change. By trying to address these conditions the tool is likely to be more acceptable to trainers and therefore it is likely to have greater face validity to trainers. It is also important to ensure that it has face validity for trainees also; this may include not containing items that they feel uncomfortable responding to. The concept of face validity is not explicit within the 'Standards'.

Another concept not covered by the 'Standards' is that of ecological validity. Ecological validity refers to the similarity between test settings and the setting in which the tool would be eventually used (Cohen, Manion et al. 2007). When a tool is trialled, the more the trial settings are similar to that under which one wishes to use the tool in the future the greater the ecological validity of the trial. The advantages of greater ecological validity is not only that the evidence of validity more specific but it is also

possible to explore how the tool will actually function in reality and will highlight any potential issues with its use. The potential disadvantage is that one has to deal with the limitations of that test setting. This concept is important because as discussed validity and reliability both only refer to the results of the test rather than the test itself (Cook and Beckman 2006, Streiner and Norman 2008); if the test circumstances are changed then the results gained may change and this will affect the validity; the concept of ecological validity is therefore implicit in the concept of validity.

2.6.3 Suitability of surgical tools

Although there are similarities in surgery to endoscopy particularly in relation to attributes required to teach a skill there are also differences. In surgery a patient is often fully anaesthetised whereas in endoscopy the patient is often awake or only lightly sedated which may alter the teaching environment and therefore the way in which the trainer must teach. The content of these tools also all differ from each other in terms of the individual items and the different aspects of surgical teaching that these cover. As endoscopy training only occurs within the endoscopy unit the items that refer to teaching in outpatients or on the ward are not relevant. Those that concentrate on teaching that occurs within the operating room are therefore more similar to teaching within the endoscopy unit but seemed to concentrate only on technical skills. This may be because the patient within the operating room is often fully anaesthetised which changes the teaching episode.

This variation in the actual content of the tool is similar to that found by Fluit et al (2010) who found the tools appeared to focus heavily on some areas of clinical teaching and not others. Fluit used domains that were taken from the CanMEDS to represent areas that a clinical teacher should role model and derived domains that they felt described attributes of the clinical teacher. None of the tools they reviewed matched exactly to these domains but they were able to comment upon this as they had explicated their concept of the clinical teacher. I have examined the surgical tools here and though they present a variety of evidence for validity it is not clear whether they measure, by the items that they use, the construct of effective endoscopy teaching. This further demonstrates the advice of Streiner and Norman (2008) discussed at the beginning of this chapter that it is important In order to judge evaluation tools that one must have a set of criteria by which to do this. The American standards enabled an assessment of the validity of the tool however just looking at the

tools with no defined construct of endoscopy teaching it is not possible to identify if these tools cover aspects of training that apply to endoscopy teaching. It is clearly necessary to understand what attributes an endoscopy trainer is expected to possess in a similar way that Fluit et al used the CanMEDS criteria to consider clinical teachers.

Essentially in order to create an evaluation or assessment tool it is important to know what it needs to measure; this is commonly referred to as the blueprint (Rust and Golombok 2009). Fluit et al (2010) had essentially created a blueprint for clinical teaching by the use of the CanMEDS competencies. The development of such a blueprint and the method by which it is developed is the focus of content validity. Rather than using this blueprint to create an evaluation tool Fluit et al (2010) used it as a blueprint on which to judge other tools. Essentially they were saying that these were the characteristics that describe effective clinical teachers which then enabled them to make a judgment of whether the tools they reviewed matched to these criteria. When considering the content of the surgical tools and the JAG endoscopy trainer tool I had not defined the attributes of an endoscopy trainer which makes it difficult to judge the content validity in reference to endoscopy trainers. Snell et al (2000) argue that instruments tend not to be generalizable as institutions have their own culture of teaching and that should be reflected in the tool. Rather than try to determine a blueprint in order to judge the suitability of other tools I therefore opted to use a blueprint to develop a new tool; this blueprint is discussed in the next chapter. The review of the above tools remains helpful in considering in what ways validity evidence can be sought.

To therefore expand on my aim at the end of chapter one the aim of this study is to create an evaluation tool that can be used to give feedback to an endoscopy trainer by either a trainee, a trainer as a self-assessment and a peer. This will be performed with reference to the standards of validity including evidence for the content, response process, internal structure and relation to other variables. Within the time frame of this study there is not the capacity to consider evidence for consequences but how this might occur will be discussed as part of the conclusion.

In the next chapter I will begin to develop the toolkit giving particular consideration to its content validity.

Chapter 3. Gaining Content Validity

In the last chapter I discussed in terms of the content of the tool how important it is to acknowledge what attributes define a high quality trainer of endoscopy in order to ensure that the content of the tool reflects this. In this chapter I will discuss how the work of Wells (2010) was used to inform the content of the toolkit and summarise his work. I then go on to check understanding of the content he defined using cognitive interviewing.

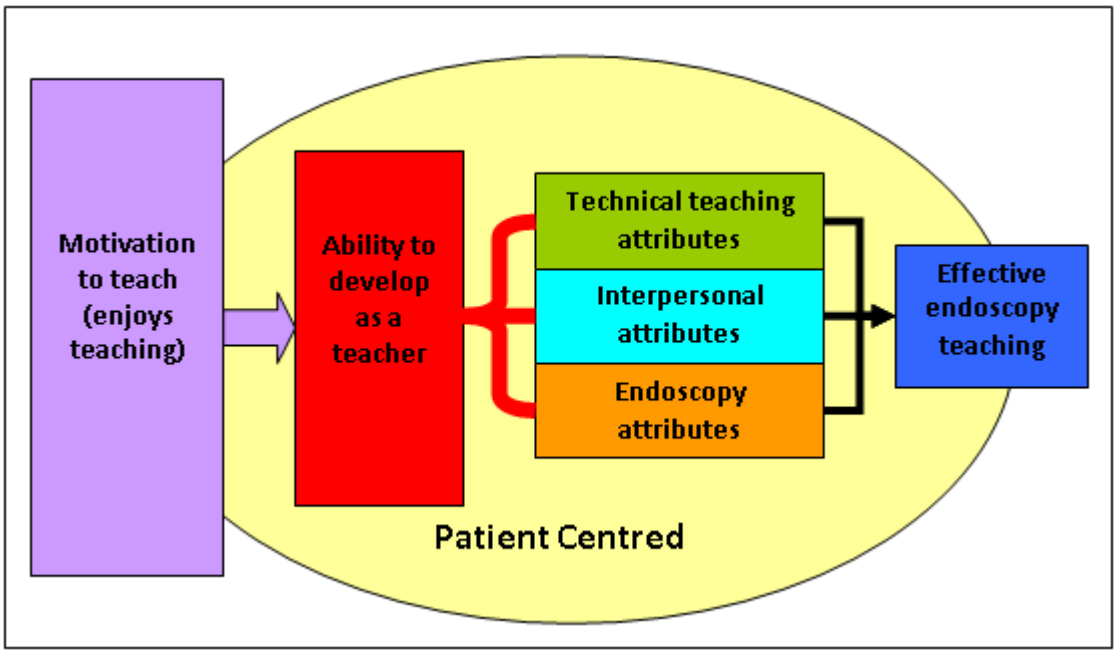
3.1 The High Quality Endoscopy Trainer

As mentioned towards the end of the last chapter the concept of providing evidence for content as part of evidence for the validity of the toolkit is two-fold. This requires a consideration of which stakeholders were involved within the development of the tool and also the items themselves. In order to create an evaluation or assessment tool it is important to know what it needs to measure; the blueprint (Rust and Golombok 2009). In considering what should be measured with regards to endoscopy training, there are several expert opinions on what characteristics an endoscopy trainer should have but clearly these are the opinions of only one individual. A group consensus regarding endoscopy also exists but this considers training as one small subsection (Teague, Soehendra et al. 2002). A qualitative study exists which considers the view of gastroenterology trainees and considers the learning experience of training in endoscopy (Thuraisingam, Madonald et al. 2006) but does not focus on the trainer specifically. An endoscopy trainer 'blueprint' could already be argued to exist through the work of Wells (2010). This is the only piece of work that fully explicates the concept of the endoscopy trainer and involved several different stakeholder groups in order to avoid the cultural reproduction as discussed in regards to the JETS tool.

Wells (2010) performed a qualitative interview study incorporating the views of trainees, trainers who work in base hospitals, 'expert' trainers involved in teaching on JAG approved courses and nurse endoscopists. Wells' (ibid) interviews focused on explicating the attributes that define a high-quality trainer. Through this process many different attributes were described and incorporated in a model shown in Figure 3-2. Wells' model states that the excellent endoscopy trainer must have not only the endoscopy skills but also appropriate interpersonal attributes and technical teaching

skills in order to be an effective teacher. Motivation to teach will also engender an attitude which will enable them to develop as a teacher. Wells also reminds us that ultimately a patient will always be present at any such teaching and therefore a patient centred approach should always be taken.

Figure 3-2. Schematic model of the attributes of an endoscopy trainer as suggested by Wells et al (2010)



This model demonstrates how the attributes were grouped and interact with each other. This makes each attribute more meaningful than if it were just to stand alone as part of a list. The above model increases the understanding of the attributes of an effective trainer but does not explain what an effective trainer actually does. Wells (2010) described that this effective teaching occurred through the processes of scaffolding and fading. In order to practice endoscopy independently a trainer must be competent in a number of skills; this is represented by the 'circle of competence' in Figure 3-3. When a trainee very first commences training in endoscopy they would be unable to complete a whole procedure as they have not yet reached the 'circle of competence', as they progress they gain in skills and move towards becoming competent; this is represented by the blue cone in Figure 3-3, as can be seen from this the trainee has not yet reached the circle of competence. This is in contrast to the trainer's skills on the right of the figure (represented by the brown cylinder) that are much greater than the minimum level of competence required to complete a procedure. As endoscopy is performed on 'real' patients every patient needs to have a complete procedure therefore it is necessary for the trainer to support the trainee in

such a way that the trainee is given the opportunity to do as much of the procedure as possible and then the trainer takes over. In Figure 3-3 the blue cone represents the rate of progression of a trainee if they were to learn how to perform endoscopy independently without the help of the trainer; if the trainer takes over the procedure at the limit of the trainee's competence then the rate of progression of the trainee would not be dissimilar if they were to just learn the procedure independently. The trainer however can maximise the amount that a trainee can do by a process of scaffolding; by scaffolding the trainer helps the trainee to achieve more through their interaction. This interaction could be verbal or through brief physical intervention. The interaction occurs within the trainee's learning zone denoted by the orange cone within Figure 3-4, and is supported by the trainer (brown cylinder). As the trainee becomes more experienced the trainer then fades away, offering gradually less scaffolding. Once the trainee reaches the limit of their learning zone the trainer may still need to take over the procedure but making this judgment at the right time is a key feature of understanding one's trainee.

Figure 3-3. The competencies of an endoscopy trainer and trainee over time

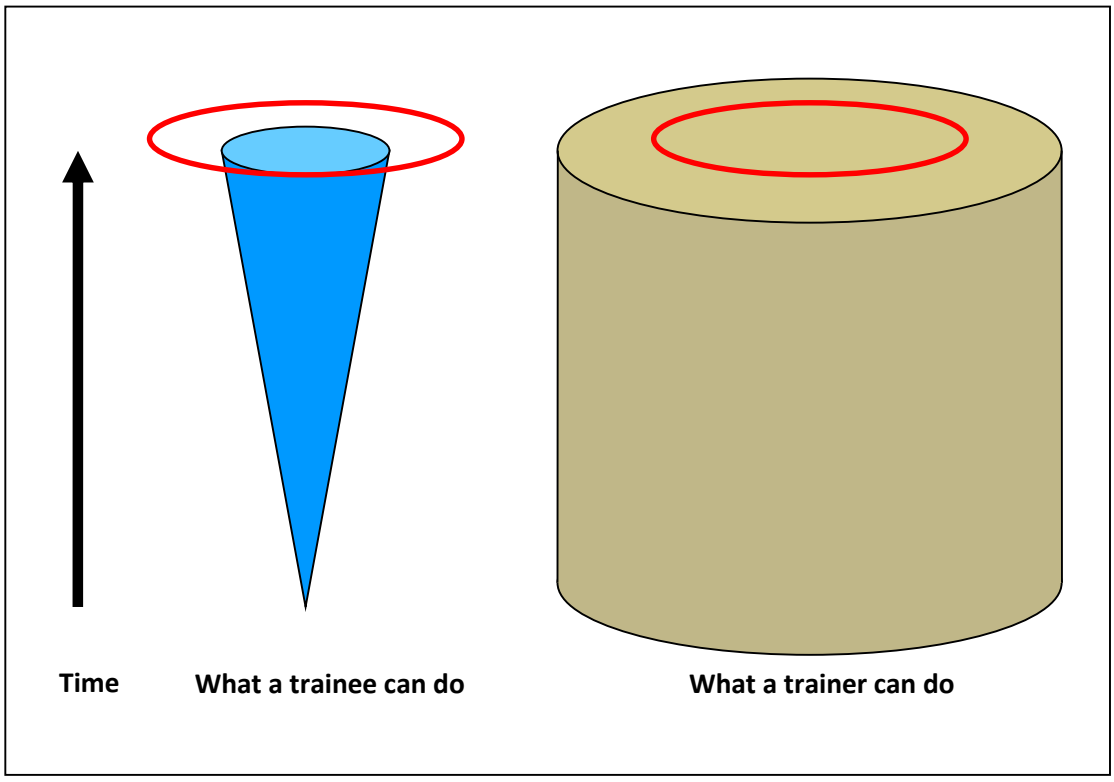
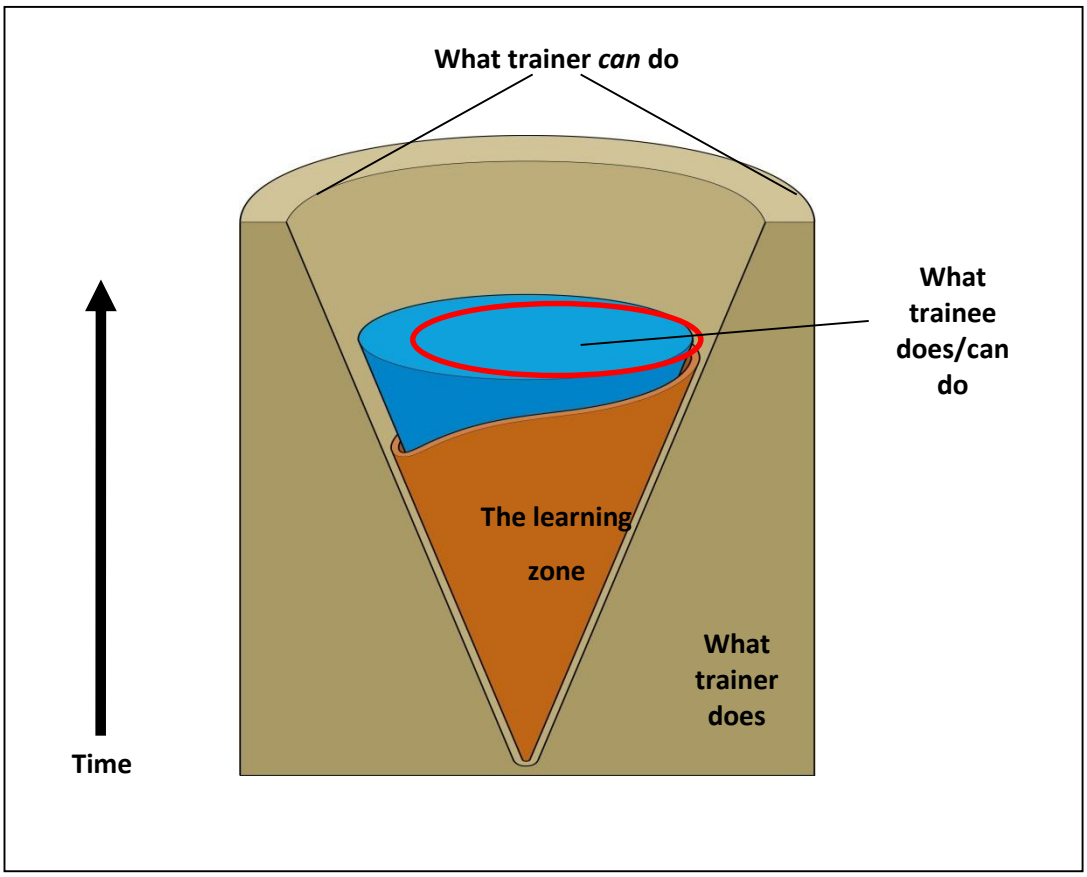


Figure 3-4. Schematic model of how a trainer facilitates trainee's learning



This model of scaffolding and fading demonstrated in Figure 3-4 gives greater understanding to the attributes described in Figure 3-2. The interaction that occurs in the orange learning zone occurs by using those attributes described in the technical teaching domain; using the attributes of technical teaching can provide the necessary methods by which to scaffold the trainee. The interpersonal skills are the 'glue' that adhere the trainee to the learning zone. The endoscopy attributes are those shown in the brown cylinder that enables the trainer to intervene appropriately and complete the procedure if necessary.

This model arose from the descriptions of training provided within his interviews although it is also described within established theories of learning. The learning zone mirrors Vygotsky's zone of proximal development (Vygotsky 1978). This is described as the gap between what a learner can do alone and what they can do with the help of the teacher. The process of scaffolding and fading has also previously been well described in the educational literature(Wood, Bruner et al. 1976).

The list of attributes described by the model in Figure 3-2 can act as a blueprint for endoscopy trainers; this blueprint can be utilized to form the basis of the toolkit. This provides strong evidence for the content validity due to the fact that this is the combined view of several different stakeholder groups. Using Wells' model to create an endoscopy evaluation tool means that the resulting tool is more likely to acknowledge the 'culture or philosophy of an organization(pg. 863)' (Snell, Tallett et al. 2000) as it is derived entirely from that setting . This is also seen in the reviews of clinical teachers (Beckman, Ghosh et al. 2004, Fluit, Bolhuis et al. 2010) each tool varied in terms of its content therefore it is important to have a template by which to understand the necessary attributes that are required to teach each discipline.

3.2 Using Wells' attributes to create a new tool

Wells (2010) postulated that his list of attributes could be utilised to create an evaluation tool for endoscopy trainers, however he acknowledged that although the above model was helpful in explaining how the attributes identified fit together to define the high-quality trainer it did not create a helpful format against which a trainer could be compared. Wells therefore re-grouped the items, largely using categories related to the Cognitive Apprenticeship model of teaching (Collins, Brown et al. 1989). This is a model that describes a theory of teaching which is based on an apprenticeship model by which practical skills were traditionally taught but expands this concept to the teaching of more cognitive processes. It emphasises the importance that these processes should be taught in the context in which they will be utilised in the future, which already occurs in endoscopy and that the role of the teacher is to model these behaviours and processes but in such a way as to make them visible to the learner. The cognitive apprenticeship model suggests a number of characteristics that create the appropriate learning environment in order for learning to take place. Wells felt that some of these characteristics could be utilised to further group the attributes that defined a high-quality trainer; these included:

- **Modelling** which refers to the trainer carrying out the task but in such a way that the learner can understand the processes both physical and cognitive that are occurring
- **Coaching** refers to the observation of the training with appropriate intervention in order to move them towards expert performance

- **Scaffolding** refers to the process where the trainer gives the trainee appropriate support in order to complete the task but as the trainee gains expertise the trainer provides less and less support
- **Articulation** refers to asking the trainee to articulate their own knowledge, reasoning or decision making
- **Exploration** involves trying to force the trainee into a problem solving mode of their own
- **Reflection and feedback**, Collins et al (1989) refer to this just as reflection and describe it as enabling trainees to compare their own problem solving technique to those of others. Wells (2010) also included feedback as part of this category
- **Content** this includes the knowledge of that specialty but does also include aspects such as heuristic strategies which Wells (ibid) opted to include as a separate category
- **Heuristic strategies** are the 'tricks of the trade' that experts use
- **Sequence of teaching** refers to the trainer increasing the complexity and diversity of tasks.

Not all of the attributes identified by Wells (IBID) within his interviews could be explained by the Cognitive apprenticeship model therefore three further categories were added. One of these was the 'learning atmosphere' which has also been added to the Cognitive apprenticeship model in other studies of clinical medicine (Stalmeijer, Dolmans et al. 2009). A further category added was that of 'preparation' as this arose from the interviews as an important category specific to the teaching of endoscopy. Finally a global category was added which included many of the important interpersonal attributes that are seen in Figure 3-2 and are not acknowledged in the Cognitive Apprenticeship model.

Wells (2010) identified 135 attributes that defined the high-quality trainer, however not all of these were measurable and also there was considerable overlap between different attributes as initially he was trying to ensure breadth within the study. He and his research team therefore through discussion and group consensus removed items that were not measurable; this was defined as not being observable by another individual; and amalgamated items 'that could be removed as the contained concept was implicit in other items – i.e. combine items whilst maintaining breadth'(Wells 2008

pg.317). In this way the number of items that were measurable was reduced to 88; these can be found in Appendix 1. Items that were not deemed to be measurable or observable were not discarded entirely as it was acknowledged that these still represented important attributes of an endoscopy trainer, rather it was suggested that they could be contained within a teaching portfolio or could form part of a handbook for endoscopy trainers. An example of an item that it was suggested could be used in a teaching portfolio referred to having attended a JAG Training the Trainers course.

In reviewing whether the attributes were measurable Wells (2008) also noted that although many of the attributes could be evaluated in a single session, some of the attributes related to characteristics that the trainer would display over a longer time period. Examples of such attributes were developing a good working relationship with a trainee or gradually increasing the difficulty of the tasks set for the trainee; these are attributes that are more likely to be displayed over the period of a rotation. Wells therefore proposed that rather than one evaluation tool being utilised a toolkit should be created with two components; these would be:

- DOTS (direct observation of teaching skills) instrument to evaluate the teaching delivered over a single session, procedure or list
- LETS (long-term evaluation of teaching skills) instrument to evaluate teaching over a clinical attachment to a hospital or rotation

Looking at when an evaluation occurs i.e. whether it is done in a single session or over a rotation has not been studied in the literature and although the reviews of clinical teacher instruments studied instruments that were to be used at either a single session or over a rotation no comment is made about these differences. One could argue that all skills seen in a single session would also be seen over a rotation and therefore it would be possible just to create an evaluation tool for just a rotation, however there would be clear disadvantages to this. One disadvantage is that a tool that examined trainers over a rotation could not be utilised by a peer and one of the aims of this project was to create just such a tool. Also in that I am trying to create a tool for formative use if a tool were to generalise over a rotation then detail would be lost and it is this detail which can be important in providing formative feedback (Harlen and James 1997). Alternatively if just a tool to evaluate a single session was created then there would be a whole range of attributes that were not evaluated as they refer to attributes shown over a longer time period or are attributes that are desirable but

would not be expected on every list, exclusion of such attributes would decrease the content validity of the tool as it would not fully describe the trainer; therefore I have opted, as Wells suggested, to create an evaluation toolkit. There is also evidence that the length of the reference period can alter a participant's response(Schwarz and Oyserman 2001). The longer the reference period can lead to under-reporting of minor events and the over-reporting of major events. This suggests that if the aim is to gain maximum detail then shorter reference periods should be utilised however this can lead to over-reporting of minor events (Schwarz and Oyserman 2001); hopefully this will be overcome by the use of both tools.

Wells provisionally allocated the items to the DOTS and LETS tools however this still contains 88 attributes making a lengthy toolkit. Although there is no optimum length of tool, level of detail needs to be balanced with the time required to complete the tool. In trying to determine whether all these items should be included in the final toolkit I felt it was important to continue to include the views of those within the field of endoscopy. The division between the DOTS and LETS was performed by Wells and his supervisory team and therefore represents their opinion only of what should be measured on every list or over a rotation. I felt that it would be relevant and important to involve stakeholders in determining which items should be included in the DOTS and the LETS. This would add further to the content validity of the tool but also, as Snell et al. (2000) highlight, teachers are more likely to 'buy in' to the evaluation if they have been involved in its development. Due to this I believed that it was important to involve the views of others in deciding which items should be allocated to each component of the toolkit and therefore needed to try and do so by gaining group consensus. Using group consensus techniques will also further add to evidence for content validity of the tool by continuing to involve different stakeholders in the tool development. It may also result in the reduction of some items helping to make the list more user-friendly.

3.3 Refining the items

Having decided to utilise the 88 attributes that define a high-quality trainer as defined by Wells (2010) I needed to examine these attributes further as they would form the items on the toolkit (and as such will subsequently be referred to as items). Streiner and Norman (2008) describe the first step in selecting items is to 'eliminate any items that are ambiguous or incomprehensible' (p77); however this poses a problem as

removing attributes based solely on their wording may result in an important attribute being lost. I therefore decided that before trying to gain consensus opinion on which items to include in each component of the toolkit I wanted to ensure item clarity. One method commonly used to examine item clarity is cognitive interviewing.

3.4 Cognitive Interviewing

Surveys and questionnaires, including evaluation tools, are often used to depict the character or opinion of large and diverse populations (Miller 2003) however they are based on the assumption that the items used are interpreted similarly by all respondents which may not necessarily be true. There is also an assumption that respondents are able to understand the item and are willing to provide an answer (Wildly and Clarke 2009). For a respondent to answer a question they need to understand the question; retrieve the relevant information from autobiographical memory; use heuristic and other decision making processes to estimate an answer and then formulate a response (Tourangeau cited in (Murtagh 2007); if this process falters at any stage this will affect the final response. For instance, miscomprehension of an item will then lead to retrieval of different information and a different final response given; this then affects the validity of the item. In particular reference to evaluation tools, not only does the respondent interpret the items but also the trainer receiving the feedback from the evaluation tool and this may differ from how the trainees interpret the items. Pilots of questionnaires may highlight some of these problems but provide no information regarding which stage of the process, understanding, retrieval, decision making or formulation of the response, has been implicated, nor will they gave any insight into how those receiving the results of the evaluation will interpret the item; an alternative to piloting a survey is cognitive interviewing.

Cognitive interviewing is a mechanism to try and improve the validity of surveys (Wildly and Clarke 2009) by asking whether an item conveys the meaning of the construct under investigation (Wildly and Clarke 2009). It is often used for questionnaires designed as health measurement scales(Carbone, Campbell et al. 2002) but can also be used in educational research questionnaires (Billings-Gagliardi, Barrett et al. 2004) and to assess the comprehension of health statements (Carbone, Campbell et al. 2002). There are several methods by which cognitive interviewing can be performed but essentially it aims to provide insight into the processes respondents use to answer

questions and highlight potential problems. It also seeks to assess whether respondents' answers represent what the question intended (Miller 2003).

It can be performed concurrently whilst respondents complete the questionnaire for the first time or retrospectively (Murtagh 2007). There are two main approaches either 'think aloud' or 'probing' (Drennan 2003). 'Think aloud' asks the interviewee to verbalise their thought processes whilst completing the questionnaire whereas 'probing' requires the researcher to ask about the respondent's interpretation of the items and to comment on the wording of the items. 'Think aloud' therefore aims to create transparency into the information a respondent retrieves and considers when formulating an answer; whereas probing has the advantage that the respondents can offer alternative wording to the questions which they believe would make the item easier to interpret or more meaningful. In practice researchers often use a combination of the two methods (Drennan 2003, Murtagh 2007).

Whilst cognitive interviewing has mainly been used in the development of questionnaires it can also be used to ensure that statements are worded appropriately. Carbone (Carbone, Campbell et al. 2002) used cognitive interviewing not only to develop a nutrition survey but also to ensure that the meaning of key nutrition messages were appropriately worded and correctly interpreted by the population for which they were intended. In order to do this a technique called paraphrasing can be used; in this strategy participants are asked to repeat the message but using their own words in order to check understanding.

3.4.1 Criticisms of Cognitive Interviews

Whilst cognitive interviewing aims to ensure that the responses given answer the question intended there are criticisms of the method; these are summarised by Drennan (Drennan 2003). The main criticism is that participants may not be able to perform the process of thinking aloud; they may find it difficult to articulate their thought processes and may appear to be struggling with the concept of the question when in fact they may be able to answer the question but not articulate it. The false environment of the cognitive interview may affect the respondent's answers. There will be an interviewer present during the interview which may not be the case when the questionnaire is actually utilised; this may create a distraction. The presence of another person may in fact cause the participant to read the question more thoroughly than they would normally do so. This artificiality can be remediated by performing the

process retrospectively once the participant has already completed the questionnaire as they would normally do so.

The final criticism regards the data analysis phase which has been criticised as lacking objectivity and is often based on the analyst's own impressions. Taxonomies(Conrad and Blair 1996) have been developed to try and improve objectivity and allow standardisation in which problems are categorised into different groups such as comprehension difficulties, judgement difficulties and so on but ultimately there will always be a degree of subjectivity.

3.4.2 Alternatives to cognitive interviews

Having acknowledged that cognitive interviews are not without fault as a methodology it is worthwhile exploring the alternatives; these include expert cognitive review, cognitive task analysis or focus groups. Expert cognitive review requires an expert to review the questions and look for any issues that might cause respondent problems; a coding system can be used to help identify and group potential problems or it can be performed without any such system (Jobe 2003). The obvious issue with this method is that respondent's difficulties with questions are often unpredictable and may not subscribe to a predetermined coding structure. A further alternative is cognitive task analysis; this again involves the use of an expert to study the items and identify the tasks a respondent would be required to carry out in order to complete the task; the expert can then decide the capabilities and limitations a respondent would likely have in completing this task (Jobe 2003). Focus groups can also be used to try and identify potential problems with questionnaire items; these small discussion groups can be used to ensure appropriate terminology and also to consider whether respondents are likely to respond to sensitive questions. However if only experts had been chosen for this study then they may have interpreted the questions in light of what they already know about endoscopy training and therefore interpret the statements differently to our non-expert target users. Focus groups could consist of experts and non-experts but are costly to organise and it is often practically difficult to get all the desired participants to be in the same place. Cognitive interviewing allows a wide range of individuals to be consulted and as each one is done separately it can be performed at the interviewee's convenience.

3.4.3 Cognitive Interviewing in Medical Education

Billings-Gagliardi (Billings-Gagliardi, Barrett et al. 2004) have used cognitive interviews within the field of evaluation of medical teaching. They acknowledged that high-stakes decisions are often made on the basis of student evaluations; however if students do not interpret the questions similarly to those reviewing the scores then this brings the validity of the results into question and affects their interpretability. They therefore performed 24 think-aloud interviews to try and ascertain issues that affected the students' choice of answer. Students had previously completed the evaluation form (which had been in use by the institution for the last five years) but then completed the form again with an interviewer present who had a pre-scripted set of probes that could be used to encourage the student to think aloud.

They found five main issues that called into question assumptions made by those who had set the questions and analysed the results. These included student uncertainty and misinterpretation of some education terminology, such as independent-learning, resulting in idiosyncratic definitions. Students tended to make decisions about teachers by comparing them to other faculty; this ranking system was not what the question had intended. Students also demonstrated, during the think-aloud interviews, that they considered what actions may occur as a result of their evaluations and altered their answers accordingly. Additionally students called on more factors than the question appeared to define; for instance in rating teaching effectiveness they did not just recall the teacher's actual performance but factors such as willingness to stay behind and answer questions. Additionally, when rating teachers, students mentioned lots of different characteristics they considered before deciding on a position on the rating scale. Interestingly when rating a teacher highly many students mentioned similar characteristics however when giving teachers lower ratings the student's reasoning for this was much more diverse. Students also tended to use the higher end of the scale when evaluating; however, as with characteristics, when they used the highest rating think-aloud interviews demonstrated that the students 'felt strongly and unequivocally positive' however when they used the second highest ratings their reasoning was much more heterogeneous including some responses that did not appear to be positive. This study demonstrates that cognitive interviewing can be of use in medical evaluation and although this study referred to evaluating pre-clinical courses one could extrapolate findings to clinical courses and teachers.

3.5 Methodology

3.5.1 Data collection

The study described above (Billings-Gagliardi, Barrett et al. 2004) led to a dilemma within my own research; it undoubtedly supports that it would be useful to perform cognitive interviewing but caused conflict as to when to perform it during the research process. The two options were to perform it either as the first step in the research process or once the initial evaluation forms have been created. The advantage with the latter is that this would help identify not only lexical issues but also some of those idiosyncratic decisions or factors considered when answering questions which may not be simply identified by considering the attributes without actually completing the tool. The disadvantage of performing cognitive interviewing once the evaluation tools have been created is that I am aiming to gain group consensus as to which attributes should be included on the tool and in which component. As I wish these decisions to be a result of the whole group then it is important that each statement means the same to all in the group so they are critiquing the attribute based on the same underlying construct; as this group consensus will largely determine the eventual evaluation toolkit. I therefore opted to perform cognitive interviewing prior to any further work. Knafl (Knafl, Deatrick et al. 2007) also argues that focusing on the items themselves without requesting an answer helps the interviewee concentrate on their interpretation of the item rather than considering their response and reduces participant burden.

I decided to perform six cognitive interviews; this is a similar number to other studies(Sepucha, Ozanne et al. 2007, Wildly and Clarke 2009). In choosing the candidates Drennan suggests that 'subjects are chosen that match the proposed sample'(Drennan 2003). I therefore chose those who would utilise the eventual tool including consultants who currently act as trainers and final year core medical or surgical trainees, as this group are about to embark on endoscopy training and it was important to ensure that even the newest trainee would understand the terminology used. I recruited participants from Northumbria Healthcare NHS Foundation Trust via email through our own personal knowledge. Each participant was sent an invitational email and an information sheet containing further information about the project. A

convenient time and place was arranged to meet those who agreed to be interviewed. All participants were required to sign a consent form. LAM interviewed all six interviewees. As the interviewees were not actually completing a questionnaire it was not possible to use a think aloud method therefore a mixture of paraphrasing (Carbone, Campbell et al. 2002) and probing (Drennan 2003) were used; probes were unscripted but mainly asked participants to suggest alternative wording if the meaning was unclear. The interviewees' answers were transcribed during the interview(Wildly and Clarke 2009).

As the list of statements is lengthy we wished to avoid participant fatigue therefore each interviewee only reviewed a selection of the statements although I ensured that each statement was reviewed at least twice and also that it was reviewed by both a trainer and trainee.

I performed all the cognitive interviewing, I had no formal training in this however I had read about the methodology (Drennan 2003) and its use in other studies (Murtagh 2007, Carbone, Campbell et al 2002) and also had discussions with others who had previously performed cognitive interviews as part of their research in order to gain practical advice. I also practiced the process with one of my supervisors (SC) prior to conducting the first interview. Notes were taken by the interviewer during the interview process by the interviewer.

3.5.2 Analysis

Following the cognitive interviews SC and myself reviewed each statement sequentially. Using a taxonomy during analysis is said to help objectify the process (Conrad and Blair 1996, Drennan 2003) however these taxonomies have been developed with regard to analysing data from interviews performed on questionnaires and therefore do not neatly fit our data. As summarised by Drennan (Drennan 2003) the categories tend to broadly fit into four headings understanding, retrieval, judgement and response formatting; however I was only interested in what the statement means i.e. were the participants able to understand it rather than the next three categories. Conrad and Blair (1996) offer a slightly more useful taxonomy shown in Table 3-1 below.

Table 3-1. Taxonomy of response processes (Conrad and Blair 1996)

Problem Type	Response stage		
	Understanding	Task performance	Response formatting
Lexical			
Temporal			
Logical			
Computational			
Inclusion/exclusion			

Although I was really only concerned with the 'understanding' response stage and therefore would have limited input into the second two columns. Conrad and Blair give detailed descriptions of each type of each problem but to summarise:

Lexical problems refer to difficulties with the core meaning of words or subtleties in reference to their scope.

Temporal problems are those where difficulties are encountered in understanding time frames to which questions refer.

Logical problems refer to problems with the use of words such as 'and' and 'or'; the use of false presuppositions and contradictions

Computational problems tend to be issues in processing and manipulating information but Conrad and Blair also suggest that any residual problems that do not fit in any of the above categories are also included in this problem type.

Inclusion/exclusion problems refer to problems relating to the scope of the question.

Although these problem types are more useful than the previous headings they still have limited applicability; for instance time frames within the statements have been left deliberately vague as attaching timeframes will form part of the next stage of the research; however I utilised it as a method of giving examples to the types of issues that arose as a result of the cognitive interviews and in considering solutions.

As mentioned in the previous chapter in order to reduce the number of items to 88 they had already been reviewed by Wells (Wells 2008): statements that conveyed the

same meaning were combined and wording used to ensure that they were measurable. In order to ensure these resultant statements still conveyed the same meaning I matched each statement with its original; these original statements often contained more detail. I then compared each interviewee's answer to both the statement they had reviewed and the original statement in order to ensure the correct meaning had been conveyed. If the meanings did not match or there was uncertainty we returned to the original interview data which was managed using N-Vivo (version 2 QSR International), a computer data analysis software package which enables the user to group data within nodes. It was therefore possible to allocate the node descriptor used to construct the original statement. I then looked at the excerpts of the transcripts within the node to ensure that the detail given in the original interviews matched the meaning given by the cognitive interviewees. If the meaning differed the statement was amended to better convey the original meaning.

This process of ensuring that the correct meaning is conveyed is referred to by Conrad and Blair(Conrad and Blair 1996) as 'author intent'. They acknowledge that often those analysing cognitive interviews are not those that originally wrote the questionnaire and therefore it is helpful to elucidate the rationale behind each question and how the author intends the respondent to interpret the question as this will help issues be more appropriately identified. They suggest a process for gaining understanding into the author's intent but as I had access to the raw data behind each statement I opted to utilise this data. Statements were also reworded at the interviewees' suggestion to increase their readability (Knafl, Deatrick et al. 2007); examples of this are given in the results section below.

3.6 Results

All participants who were approached agreed to participate; this included one surgical and two gastroenterology trainers; one was female; one surgical and two medical trainees; two were female, one male. I met all interviewees at their place of work; for the trainers this was in their office, whilst for the trainees this was in a room in the education centre of their place of work. Each interview lasted between 40 minutes and one hour.

Following cognitive interviewing each item was reviewed item-by-item; initially item analysis was carried out by myself and SC; with the results of our analysis along with

the results of the cognitive interviews sent to my other supervisors for review and further amendments. Thirty-five statements in total were amended following this process. The interviewees answers were compared with each other to see if both trainees and trainers interpreted the statement in the same way; these answers were then compared with the original statement; if there was still any uncertainty with regards to the meaning of the statement we returned to the original qualitative data and reviewed the original transcripts to ensure the meaning had been correctly captured. The final list of amended statements is shown in appendix 2.

3.7 Discussion

3.7.1 Lexical problems

Using Conrad and Blair's taxonomy the large majority of problems identified were lexical problems; this was not unexpected as participants were asked to paraphrase which is most likely to show up issues with the meaning of items rather than any of the other problem types. One problem was with the meaning of words within the context of the statement; for instance in the item

'Everyone's roles with respect to training were clear'

There was uncertainty as to the meaning of the word 'role' as to whether it referred to who everyone was or how they were to act within the training process; comparing the above statement with the original statement it was more apparent that the latter interpretation was the one intended by the author. In order to make this more transparent to the respondent the item was reverted back to its original version within Wells' work;

'The trainer clarifies everyone's roles before a training encounter so that each individual knows how they are involved in the training process.'

A further lexical issue was that of understanding the meaning of the phrase 'whole process' in the item

'The trainer taught the whole process of endoscopy to the trainee'

One of the interviewees interpreted this as meaning the non-technical aspects of endoscopy whilst another interviewee initially suggested this referred to 'all the surrounding bits of endoscopy such as post-procedure care' but then doubted themselves and wondered if it meant that the trainer taught all the practical components of endoscopy and did not just assume the trainee already had knowledge or experience of those components. Again I referred back to the relevant N-VIVO node and reviewed the excerpts that supported this item and discovered that it referred to what the latter interviewee had first surmised; in order to clarify this I therefore added the examples referred to in the N-VIVO excerpts so that the item read

> *'The trainer taught the whole process of endoscopy to the trainee e.g. the indications, consent and communication'*

A further example of when examples were added was in the item

> *'The trainer ensured that the trainee was physically comfortable'*

The interviewees interpreted this statement correctly but professed doubt as to whether in fact the item was referring to something else; therefore examples which had been included in the original item were added to help clarify the meaning

> *'The trainer ensured that the trainee was physically comfortable (including neither being too tired nor in actual physical discomfort)'*

Above are two illustrations when adding examples can help clarify the meaning and thus overcome lexical issues; however during the cognitive interviewing it was also found the presence of examples can also confuse respondents; for instance in the statement

> *'The trainer used other equipment that can support teaching (e.g. the magnetic imager, models etc) appropriately'*

This confused one of the interviewees although he correctly interpreted the item he professed he was unsure what teaching aids could be included. He gave as an example that he had often seen whiteboards and diagrams used to support teaching effectively but as the examples were more high-tech than this he would not know whether this statement was meant to pertain to high-tech aids only. During the analysis phase I

reviewed this phrase in N-VIVO and found that both low and high-tech aids had been discussed I therefore changed the examples to cover both elements;

> *'The trainer used teaching aids that can support learning (e.g. the magnetic imager, diagrams, models etc)'*

Some of the items were altered to increase their readability; although the interviewees interpreted the meaning correctly they felt it could be worded better; this does not fit neatly into any of the problem types but sits most closely with lexical problems. An example of this is the item

> *'The list was populated with cases appropriate to the needs of the trainee (in terms of volume and nature of cases)'*

Although the interviewees correctly interpreted the meaning of the item they felt it could be written more simply and therefore was re-worded using one of the interviewee's suggested wording; this also meant the statement was more measurable. The resulting item read

> *'A trainer prepares the endoscopy training lists to meet the current needs of his trainee both in volume and the nature of the cases on the list'*

3.7.2 Logical problems

Several of the logical problems involved highlighting potential contradictions either within the item itself or with other statements. An example of a perceived contradiction within an item was

> *'The trainer provided continued supervision for his former trainee even when the trainee was fully trained'*

Although the trainer correctly interpreted its meaning, possibly because he had previously been in this position, the trainees did not. They thought that someone who was fully trained would no longer need to be supervised and therefore a contradiction; when I reviewed the original data it was appreciated that the word 'supervision' implied continuing observation and teaching, and that it was the norm for trainees to continue to gain support even when appropriate competencies had been achieved; the item was therefore amended to

'The trainer provided continued support for his former trainee even when the trainee has achieved competence to 'sign-off''

An example of a perceived contradiction between statements can be found when comparing the item

'The trainer adjusted the position he was standing in the room appropriately, withdrawing as the trainee progressed'

And the item

'The trainer closely observed the process and was aware of what the trainee was doing'

The meaning of the first of these two items was interpreted appropriately by all interviewees and matched with the original excerpts that supported this item which described that sometimes the trainer might withdraw outside the room. One of the interviewees when reviewing the second item questioned how could the trainer closely observe if it is accepted, as in the last statement that he had physically withdrawn himself; this does indeed appear to be a contradiction, As a result of this discrepancy I reviewed this item in the original N-VIVO data and found the data supporting this from the original interviews did differ for instance one said that the trainer should always be in the room whilst another supporting statement was about the trainer being aware of the whole process including pre and post care which he may not need to be in the room for. These differing views are likely to have arisen due to the fact several different people were interviewed by Wells (Wells 2008) and their opinions on some subjects may differ. Both of these statements have therefore gone forward unchanged to the next stage of this process so that I can try and gain a consensus opinion as to what the latter statement means rather than imposing a personal opinion on it.

Conrad and Blair(Conrad and Blair 1996) also argue that repetition between questions is a logical problem as it leaves the respondent baffled as to why he (or she) is being asked the same question twice. Although not picked up by our interviewees, though this may have been because each interviewee did not review every statement, during our analysis I noted that the below two statements were measuring the same underlying construct

'The trainer agreed rules for teaching with the trainee'

And

'The trainer agreed the rules of the training and was consistent in the application of these rules'

Therefore I removed the first of these two statements. The reason I chose this second item was that it not only included the concept in the first item but was more wide reaching in that it would remind trainers that they should also apply these rules. The second item could be criticised in that it actually evaluates two concepts and therefore if a trainer was scored poorly on this item it may be due to the fact that they did not set the rules or was not consistent with said rules, however I felt that if a trainer were to set rules but not follow them this was no better than not setting rules in the first place and therefore felt it was acceptable that this item contained two concepts.

3.7.3 Computational problems

Computational problems as mentioned earlier are often issues with memory recall but can include anything else which does not fit into one of the other problem types. An example of a problem of memory retrieval was in the item

'The trainer agreed SMART goals for the session with his trainee at the start of the list'

Although all interviewees recognised the acronym 'SMART' and recall some of component words no interviewee could recall the exact definition and therefore the trainer may not be evaluated on all parts, the definition has therefore been added to the acronym,

'The trainer agreed SMART goals for the session with his trainee at the start of the list (S = specific, M = measurable, A = achievable, R = relevant, T = can be achieved in the timeframe)'

A further issue that arose from the cognitive interviews involved the statements

'The trainer taught the trainee to communicate appropriately with the nurses'

'The trainer taught the trainee to communicate appropriately with the patient'

Although these statements were correctly interpreted there was uncertainty regarding the word 'taught' and whether these statements represented skills that could be

encouraged but not taught; interestingly this word was volunteered by several of the interviewees and was in fact the term used in the original statements and therefore in both of these items the word 'taught' was replaced with the word 'encouraged'. This can be defined as a computational error as it does not fit in with the memories of either teaching or being taught.

3.7.4 Temporal problems

No temporal problems were identified during analysis of the cognitive interviews.

3.7.5 Author intent

Along with the above problems identified, a further issue emerged, which relates to author intent. Whilst matching the items to the original data, I occasionally found that the original statement contained a slightly different or extra implied meaning. Even though the interviewees had correctly interpreted the statement they had in fact missed an element of the original statement because it was not included in the statement they reviewed. An example of this occurs with the item

> *'The trainer always ensured that the patient was comfortable and safe'*

The original statement in fact reads

> *'The trainer's prime concern is always that the patient is comfortable, safe and has his (or her) dignity maintained throughout the training'*

Because it is not mentioned in the statement they were reviewing no interviewee mentioned the concept of dignity in their answers. This is clearly the author's intent as it is mentioned in the original statement. Following cognitive interviewing and despite its correct interpretation I altered the statement to ensure the concept of dignity was included, so the item read

> *'The trainer always ensured that the patient was comfortable and safe and their dignity was maintained'*

and was included in the next stage of toolkit development

3.8 Conclusion

Cognitive interviewing has highlighted potential problems for mis-interpretation of the construct under investigation. These problems left unaltered may have meant that decisions about an item's inclusion would have been based on different interpretations by different members of the panel. They also have the potential to reduce the content validity of the tool as those that utilise the eventual tool may have also interpreted items differently.

In conducting the interviews at this stage it meant that I could only ask participants to paraphrase the items rather than think aloud an answer, which would have been possible to do had they been able to complete the evaluation tool. At times participants found it difficult to paraphrase, often because for some items they felt that the meaning was so obvious it was difficult for them to find alternative words. Asking them to do this may have added another cognitive task and this task may have changed the outcome because they had difficulty finding alternative wording and I then attached meaning to the choice of these alternative words. The advantage of conducting it at this stage was that it conferred an advantage within the next stage of the toolkit development as variability in interpretation of items may have altered the decisions of those in the consensus process (discussed in the next chapter).

Note taking occurred during the interview in order to be cost effective but obviously has disadvantages. It may have impeded the interview in terms of timing and may not have been word accurate although attempts were made to ensure that the transcriptions were as accurate as possible in terms of documenting the wording used by interviewees. An alternative could have been to have an observer in the room either performing all the transcribing in order not to disrupt the flow of the interview or concurrently which would have improved reliability as the transcriptions could have then been compared. A further alternative would have been to record the interviews for later transcription.

It was useful to perform the cognitive interviewing on both trainees and trainers as, has been shown by some of the examples above; they interpreted some of the items differently. Not only might this have influenced responses but trainers might have received a different message from the feedback than that intended.

Using a taxonomy (Conrad and Blair 1996) to categorise types of problems for response type and the understanding element of the response stage helped illustrate the type of

problems that were encountered even though two elements, task performance and response formatting were not used. This was helpful when determining what changes to the items should be made; it enabled me to pinpoint the reasons why different respondents interpreted statements differently and aided decision making in terms of making changes to the items.

One of the criticisms of cognitive interviewing mentioned above is that it is subjective. Having completed the process I believe that by sticking closely to the author intent made it less subjective; this was substantially aided by having access to the interview excerpts that represented each attribute as I was then able to use these to support decisions about the way items were changed. Without this data it would have been difficult at times to decide when respondents' interpretations differed, which one was the 'correct' interpretation; without the interview data I would have likely had to use my own interpretation which would have made the process more subjective. At times it was not possible to not make a decision on what I thought was best; an example of this is seen in section 3.7.3 in terms of discussing the two items that refer to the ground rules as the item I chose was the one I ultimately felt was best. The subjectivity of the process was decreased by conducting the process with one of my supervisors and then the changes reviewed by two other supervisors.

There is an argument that I should now repeat the cognitive interviewing process following the changes I have made; however I did not want to pre-empt subsequent potential changes. I did not regard these statements as the finished article; instead they provided the starting point for improving and selecting statements. Cognitive interviewing has had flaws as discussed above but has ultimately contributed to validity evidence within the category of Response Process; as discussed in Chapter two this is often an overlooked area of validity and in fact both reviews comment on the lack of evidence that the tools they review provide for the response process category. Trying to ensure that the items are as clear as possible will help optimise the response process.

Following cognitive interviewing I wanted to ensure that all items would potentially be appropriate to be included in the final evaluation tool. In setting out my aim at the beginning I stated that I wanted to create a tool that could be completed by a trainee, peer or self and therefore the items needed to reflect this. I have previously mentioned that Wells had reduced the items to those that were measurable but had not

necessarily designed the toolkit to be used by peers, trainees and trainers. JRB, SC and myself therefore agreed criteria for inclusion: that the statement must be measurable by all who complete the tool including a self-evaluation, a peer evaluation and a trainee evaluation, and that the statement must be generic to any trainee regardless of the stage of training and procedure performed. SC and myself therefore applied these criteria to the list of items, which was then subsequently reviewed by JRB and MRW. The following items did not meet the agreed criteria:

'The trainer dealt with any lack of insight in the trainee'

'The trainer taught the trainee about loop resolution'

'The trainee let the trainee handle the endoscope outside of the patient before using the scope on the patient'

'The trainer provided continued supervision for his former trainee even when the trainee was fully trained'

I have therefore decided that these items should not be included in the final toolkit and have been excluded from further review. I acknowledged that these may represent important endoscopy trainer attributes however it is essential that items included in the final toolkit are measurable by all and relevant to every list or grade. In the next chapter I will discuss how the rest of the attributes were reduced and allocated to the DOTS and the LETS using a consensus technique.

Chapter 4. Group consensus

In this chapter I discuss how group consensus was used to select and further refine the items to be included on the toolkit. I initially discuss the various different group consensus techniques before opting to use the Delphi process. I then critique the Delphi process further before describing how I used it within this study and the results obtained.

4.1 Gaining group consensus

Following the cognitive interviews there were 83 items that could be included in the toolkit, however these items still needed to be allocated to either the DOTS or the LETS component of the toolkit. As I wanted to create a toolkit that would be readily accepted by the endoscopic community and used frequently I was also concerned that there may be too many items. There is limited research into the ideal tool length and in a review of clinical teacher evaluation tools the item length of the tools reviewed varied from one to 58 (Fluit, Bolhuis et al. 2010). Streiner and Norman (2008) give an overview of research into the length of mailed questionnaires, there is some evidence that shorter questionnaires have an increased response rate (Yammarino, Skinner et al. 1991) cited in Streiner and Norman 2008) however there is also evidence that by adding interesting questions that this may also increase response rate (Burchell and Marsh 1992) cited in Streiner and Norman 2008). It therefore appears that the content of the toolkit is critical to its success; it was important to ensure that the content was appropriate but pragmatically considering tool length also seemed important.

The original research performed by Wells(Wells 2008) derived the statements from twelve interviews; although he opted for 'maximum variation sampling' (Patton 1980 in Wells 2010) aiming to cover the range of different views across the endoscopic community there may still be viewpoints that have not been covered by these interviews. As a result of this when deciding which statements should remain in the final evaluation toolkit and to which tool they should be allocated I wished to gain a breadth of views essentially aiming for group consensus. In addition to allocating the attributes to the different components of the toolkit it has also previously been noted that involvement in the development of a tool can lead to later buy in(Snell, Tallett et

al. 2000). Ensuring that group consensus was gained would also contribute towards the content validity of the final toolkit.

Consensus methods are widely used in medicine as a method of synthesising information when it cannot be done by more conventional methods such as meta-analyses(Jones and Hunter 1995). They have been used in a variety of fields including that of training and education but also service development, use of technologies and appropriateness of clinical interventions (Jones and Hunter 1995) (Fink and Kosecoff 1984). Arguments for using a group to gain consensus include the view that pooled intelligence is felt to be greater than that of individuals; the judgement of more people is likely to get closer to the truth; complex ill-defined problems can only be addressed by pooled intelligence and the consequences of research is more likely to be accepted following participation by the group (Moore 1987 cited in (Clayton 1997). Criticisms of groups per se though are that often the outcome is not the perceived wisdom of the entire group but that of one or two dominant individuals in a group(Mckee, Priest et al. 1191) and that, although one would think that a group would conform to the norm or safe option, the perceived safety in the group means that the group can move towards a more extreme pole of opinion. According to Clayton (1997) this phenomenon was dubbed 'risky shift' and was seen in studies where group discussion seemed to intensify extreme attitudes, beliefs or perceptions rather than moving back towards the norm. In order to try and address these criticisms structured group techniques have been developed of which the main three are consensus development conference, nominal group technique (or the expert panel) and the Delphi technique.

The consensus development conferences were popular in the late 1970s; Fink (Fink and Kosecoff 1984) reports the U.S. National Institute of Health organised over 40 such conferences between 1977 and 1984 largely related to new technologies or treatment options. Members are selected to a panel and then the panellists hold a public meeting at which representatives involved with the topic under question are invited to speak with further questions and discussions from the floor. Following the meeting the panellists reconvene to create a consensus statement based on the evidence given at the meeting(Mckee, Priest et al. 1191).

The nominal group technique involves one or two highly-structured face-to-face meetings normally involving nine to 12 participants(Mckee, Priest et al. 1191) in which participants either create or discuss predefined statements relating to the topic in

question then each individual privately ranks the statements. The facilitator tabulates and presents these rankings; this overall ranking is then discussed by the group before individuals re-rank the statements a second time(Jones and Hunter 1995).

The Delphi technique does not require participants to meet but is conducted either by mail or electronic questionnaires with the central features being iteration and feedback(Crisp and Pelletier 1997). Participants are asked to provide opinions on the topic in question; the researcher then collates these opinions which are then reviewed by the rest of the group who rate or rank them; these are then returned to the researcher who collates the scores. A summary of the results is then presented to all participants who then re-rank the items based on the rest of the group's opinion; this process continues for a predefined number of rounds or until consensus has been deemed to be reached. The Delphi technique is said to be 'modified' when the statements under review are not derived as part of the Delphi process but originate from elsewhere (Murry and Hammons 1995).

In choosing which of these methods should be used in this project I needed to consider the advantages and disadvantages of each. The funding and organisation required to host a consensus development conference is considerable and therefore beyond the scope of this research project. The disadvantage of the nominal group technique is that it can only handle small numbers of participants in the region of nine to 12; this is a similar number on which the original interview work was performed(Wells 2008) and as the argument for using a group technique was to gain a greater breadth of opinion I required a method that could deal with larger numbers. In addition to this, the current DOTS form used by JAG (Figure 2-1) was formed by nominal group technique (personal communication from JRB) using those professing a special interest in training such as JAG course trainers. In order to include Base Unit trainers, trainees and nurse endoscopists, all of whom may use the final tool, I decided a modified Delphi technique was most appropriate as it can handle greater numbers, and avoids the difficulty of arranging for large numbers of health professionals to meet(Murry and Hammons 1995). In addition, as Delphi participants never meet and are anonymous to each other, it minimises the influence of dominant individuals (Stewart and O'Halloran 1999). This may be particularly important when using a varied group of experts as if the group were to meet some members may feel intimidated by the 'expert' trainers and therefore less likely to offer their opinion.

4.1.1 The Delphi technique

Considering the Delphi technique further; it was first introduced by the RAND corporation in the 1950s as a means of gaining group consensus from experts regarding military strategies(Murry and Hammons 1995). It was named after the oracle on the island of Delphi who was able to accurately predict the future(Villiers, Villiers et al. 2005). The Delphi technique since its inception has been used by a variety of industries and organisations for a variety of different purposes which has required an ever changing definition of what it entails(Crisp and Pelletier 1997). In response to these changing definitions and uses there has been an attempt to classify the Delphi technique into sub-types; again these descriptions differ but Linstone and Turoff (1975 cited in (Villiers, Villiers et al. 2005) suggest the following descriptors:

- Conventional Delphi involves the prioritisation of facts. A questionnaire is sent out to experts with a second questionnaire then sent based on the results of the first. Each successive round is accompanied by feedback from the previous round with the aim of gaining consensus on the accuracy of the facts or to gauge support.
- Real-time Delphi is similar to the above but occurs over the course of a meeting
- Policy Delphi is less concerned with reaching a consensus but in gaining views from different experts so that multiple points of view are heard and represented.

The method I have used most closely matches to the conventional method as it was important to gain consensus and support for the final version of the tool that would be used by the respondents.

Despite these differences the defining factors that run though each type are that a Delphi should be an iterative process with participants having the opportunity to change their standpoint. The second factor is that it is an anonymous process with participants not known to each other (Crisp and Pelletier 1997) allowing no-one in the group to dominate.

4.1.1.1 Criticisms of Delphi

Two criticisms of the Delphi technique, which are also levelled at other group techniques, are what is considered to be consensus and who counts as an expert. Consensus can be divided into two areas; it can be used to assess the level of agreement (consensus measurement) or to resolve disagreement (consensus development) (Jones and Hunter 1995). The concept of consensus is one of the most contentious parts of the Delphi process with some even arguing that consensus reached in a Delphi study does not really represent true agreement (Crisp and Pelletier 1997) as even if all members of the group agree this does not mean that this is necessarily the 'truth'; 'there is a danger of deriving collective ignorance rather than wisdom'(Jones and Hunter 1995).

Whilst the majority of studies seek consensus (Stewart and O'Halloran 1999, Campbell and Cantrill 2000, Elwyn and O'Connor 2006) there is no uniform agreement as to what counts as consensus; how researchers are to know when consensus has been reached and how to represent consensus statistically. Participants are often asked to rate statements on a Likert scale(Villiers, Villiers et al. 2005, Elwyn and O'Connor 2006) whereas others ask for statements to be ranked in order of importance or relevance(Okoli and Pawlowski 2004). Stewart (Stewart and O'Halloran 1999)simplified this process even further and just asked participants to accept or reject the statements.

Following this rating or ranking process how the level of agreement and spread of opinions is displayed varies widely but is normally represented statistically. The greatest area of contention, however, is numerically where does the cut-off lie that denotes that consensus has been reached. There is no universally accepted figure and this varies from study to study. For instance Okoli(Okoli and Pawlowski 2004) and Villiers(Villiers, Villiers et al. 2005) deem that 70% agreement represents consensus. Stewart (Stewart and O'Halloran 1999) encountered an issue with consensus when performing a Delphi to consider appropriate tasks for the pre-registration year. In their study consensus had been set at 95% however they discovered that if they maintained it at this level then all laboratory and clinical investigation tasks would have been excluded; this demonstrates the pitfall of setting consensus too high. Critics argue that the Delphi technique is not a robust method as it is possible to manipulate the results by moving the threshold for consensus. Supporters of Delphi argue that this can be overcome by being explicit about the threshold for consensus prior to study

commencement. The alternative method of ascertaining that consensus has been reached is to stop when there is no longer any change in opinion between rounds i.e. 'the point of diminishing returns is reached'(Fink and Kosecoff 1984).

Deciding when to stop is also an important element in deciding if consensus has been reached. If the number of rounds is set prior to commencement of the process then the danger is that the Delphi can become an artificial exercise as it may be terminated before consensus has been reached. An alternative to this is to conduct no further rounds when there is no further change in results however this may lead to participant fatigue and increased drop-out. Taking into account both viewpoints three rounds are said to be optimal (Villiers, Villiers et al. 2005, Hsu and Sandford 2007) as most convergence of responses occurs between rounds one and two(Murry and Hammons 1995).

The second point of debate surrounding Delphi and indeed all group techniques is that they utilise the concept of expert. Although the concept of using a group to make decisions or draw conclusions is that 'several heads are better than one'; who those several heads belong to plays a large part in determining the validity of the final results. For instance builders deciding on guidelines as to who should have an MRI for back pain would be less valid than if orthopaedic surgeons were to complete the same process. The process of selecting experts is critical to Delphi and 'to authorise its validity and superiority' (Clayton 1997). Murry (Murry and Hammons 1995) argues that 'expertise implies that the individual panellists have more knowledge about the subject matter than most people'; this is in keeping with the Collins English Dictionary definition which defines an expert as 'a person who has extensive skill or knowledge in a particular field'. However what is accepted as expert varies from situation to situation, Moore (1987 cited in(Clayton 1997)) gives the example that 'a nuclear physicist is an appropriate expert if the Delphi concerns atomic energy and a resident of a neighbourhood is an expert of what a community's goals should be'. Pill (Pill 1971) also argues that in fact an expert could be defined as anyone who 'can contribute relevant inputs' and includes the example that this 'might include a consumer in the case of constructing consumer preference scales'. Sinha (Sinha, Smyth et al. 2011) also argues that one should consider the concept of expert more widely and consider who should have influence on the consensus reached; for example Sinha (ibid) argues when performing a Delphi study to decide on what outcomes should be measured from

clinical trials then both patients and clinicians should also be included as certain outcomes may be more important to patients than pure scientists and hence they should be deemed experts as well as traditional research experts.

As well as considering the degree and nature of expertise when selecting participants they should be sufficiently interested and motivated to take part in all rounds of the Delphi process(Clayton 1997). Clayton (ibid) also notes that it is important to include those who will ultimately use or act upon the results of the Delphi.

Lastly how many experts are needed? The literature is mixed about the optimal size of the panel. Cochran (1983 cited in (Murry and Hammons 1995) found that as panel size increases so does reliability and error is reduced; however it is generally recognised that few new ideas are generated once panel size exceeds thirty for a homogenous group. The suggested panel size appears to be thirty for a homogenous group and five to ten per category for a heterogeneous group(Crisp and Pelletier 1997, Villiers, Villiers et al. 2005) although there is no data to support this.

4.1.2 Methodology used in this study

As highlighted above there are no firm guidelines for how a Delphi should be conducted and certain areas can cause much debate; therefore it is suggested that researchers using this technique should be explicit about the methodological decisions they have made and justify those decisions (Sinha, Smyth et al. 2011).

4.2 Round 1

4.2.1 Methods

4.2.1.1 Recruitment

Firstly I considered who should be classified as experts; experts are selected for a purpose and not at random(Hasson 2000). Options in determining who are experts include using positional leaders, authors of relevant publications or those with first-hand relationships with the particular issue (Hsu and Sandford 2007). Positional leaders on this topic are those who sit on the JAG committee and teach on the 'training the trainer' courses as this group teaches others how to teach and therefore can be deemed to have particular skill or knowledge in this area. However the current DOTS form as previously mentioned, was created by a nominal group technique using a group similar to that described above who are likely to reproduce rather than challenge

the dominant view (Bourdieu 2004) and perpetuate existing attitudes. In order to overcome this potential risk base hospital trainers were also included as it is this group the toolkit is primarily aimed at and therefore will have first-hand relationships with the issue and will subsequently be the group to use the toolkit. The above argument can also be used in relation to trainees who will have important opinions as to what attributes their trainers should possess; other literature has also stressed the importance of including trainees in the creation of assessment tools of clinical teachers (Donner-Banzhoff, Merle et al. 2003) and there is evidence to suggest trainees place greater emphasis on different attributes to that of trainers (McLean 2001). In order to ensure all views are represented and because they enter endoscopy through a different pathway to many other endoscopy practitioners nurse endoscopists were also included. Nurse endoscopists contribute significantly to the endoscopy workforce but may have had different experiences of training to other endoscopists. They will have potentially had different previous experiences of being trained within their parent specialty compared to doctors; this may be reflected in differences in how they view training in endoscopy. Wells (2010) also selected the same four groups for his interviews and by utilising these same groups consensus in this study with a larger group of participants will help add strength to Wells (ibid) original work but will also ensure that none of the voices of all four sub groups are lost. All of these groups will also use the final toolkit making their participation in its development important (Clayton 1997).

Participants to each of these groups were recruited in the following ways:

- JAG members and 'training the trainer' course leads were sent a personalised email invitation from JRB.
- Base hospital trainers were recruited through the Northern Region Endoscopy Group (NREG)(Rees and Rutter 2010); the chairman of NREG sent an email to the NREG lead at each trust within the Northern deanery who then disseminated it to trainers within their trust.
- Trainees were recruited through a short oral presentation at a gastroenterology and surgical study day and were given a letter detailing the project with a reply slip

- Nurse endoscopists were recruited locally by contacting the endoscopy department of each trust who were then emailed directly; this was supplemented by nurses nationally through JAG.

Those who were emailed were asked to reply to indicate their interest and those on the study days were asked to return the reply slip. Adopting this technique rather than just sending the first Delphi is a method of trying to minimise attrition (Sinha, Smyth et al. 2011) and ensure that participants are motivated to participate. I aimed to recruit 10 to each group as suggested for heterogeneous groups such as these(Crisp and Pelletier 1997) but because it is important that those who participate will later be using the tool (Clayton 1997) it was important not to exclude anyone who later may have influence over the tool's application therefore all identified training leads were invited to participate.

4.2.1.2 Process

Once the recruitment process was complete all participants were sent a copy of the list of statements. In a traditional Delphi the first round normally consists of a series of open-ended questions in order to ascertain the panel's opinions (Jones and Hunter 1995) however in a modified Delphi this stage is removed and the participants are given a structured questionnaire (Murry and Hammons 1995). I opted to perform a modified Delphi as I wished to use the attributes derived from Wells' (2010) qualitative interviews as previously discussed; this also helps reduce the number of rounds (Stewart and O'Halloran 1999) and hence hopefully reduce participant fatigue. It can also help avoid the 'collective ignorance' which can be a criticism of the Delphi (Jones and Hunter 1995) as the items have been derived from empirical study and therefore have a firm theoretical basis from which subsequent decisions are made.

Each round was conducted via surveymonkey(Surveymonkey 2008); participants were sent the link via email but were able to request a paper copy if they so wished. Using email and the web has been shown to increase the speed of the process(Hsu and Sandford 2007) and to decrease non-response rates. The survey contained instructions which reminded participants of the context and aims of the research, the definitions of the DOTS and LETS and the instructions on how to complete the questionnaire. The participants were asked to initially provide some demographic data and then asked to rate each attribute on a five-point Likert scale (Likert 1952) from strongly disagree to

strongly agree as to its suitability for the DOTS and then for the LETS. A five point scale was chosen as fewer points than this reduces the reliability of the scale; too many steps increases the time taken by participants to rate each item and people are unlikely to be able to discriminate between more than seven levels (Streiner and Norman 2008). Participants were also given the opportunity to suggest any modifications or justify their answer for each attribute; a copy of the questionnaire is included in Appendix 3. The panel was also asked to review the list of statements that had not met our pre-defined inclusion criteria at the end of cognitive interviewing to see if they felt that any of these should be included. They were also given the opportunity to suggest any statements or attributes they did not feel had been covered in the other statements but should be included in the toolkit. Non-responders were sent two reminder emails at two-week intervals.

4.2.1.3 Analysis

Results were analysed both quantitatively and qualitatively. Statistically levels of agreement were initially reviewed using SPSS (SPSS 2005). Agreement was defined as a score of 4 or more on the Likert scale; and consensus was counted as 70% of participants 'agreeing' with the statement. Statements that did not meet this threshold were excluded from subsequent rounds unless from the free-text feedback the reason for rejection was due to unclear wording or meaning. In these cases the statements were reworded for clarity and included in the next round for further review.

Any statement that achieved greater than 70% was felt to have achieved consensus for inclusion; the next step was to review whether it should be included in the DOTS or LETS. For each statement the statistics for the LETS and DOTS were calculated; the percentage of participants who ranked the statements $\geqq 4$ and the percentage that ranked it 5 (strongly agree) were noted for the DOTS and then the LETS. A Wilcoxon paired test was performed for every statement to review whether the difference in scores between the LETS and DOTS was statistically significant. If the statement received greater than 70% agreement for one element of the toolkit and then had both the highest levels of total agreement and the highest number of 'strongly agree' for the same element of the toolkit i.e. either the DOTS or LETS and this difference was statistically significant then it was deemed to be accepted for that element and was not resubmitted in the next round.

Free text comments were reviewed thematically and statements adjusted accordingly, rules for managing these decisions was taken from Yeates's work shown in Table 4-1(Yeates, Stewart et al. 2008) . If the wording of any statement was changed to a degree that the research team (JRB, SC and LAM) felt its meaning could be interpreted differently then it was submitted for review in the next round independent of the statistics. If the wording was only changed minimally then decisions for inclusion in the next round was based solely on the statistics.

Table 4-1. Rules for managing panel responses (Yeates, Stewart et al 2008)

Delphi rules for managing panel responses
1. Where two or more statements express similar ideas, they may be amalgamated or re-phrased more succinctly as long as the individual concepts contained within the statements are not lost
2. If two phrases express essentially the same concept, the researcher may choose one statement over the other to express the concept (and omit the latter). The main determinates of this choice will be simplicity and clarity
3. When amalgamating two or more statements which contain a number of concepts, a concept may be removed if it is already contained in other statements elsewhere in the list
4. If the phrasing of a statement is cumbersome, it may be re-phrased more succinctly as long as the concept is expressed
5. The researcher's interpretation of the meaning of the statement (or its component parts) is the key to determining whether a case has been retained or lost. As long as the meaning is still expressed, the statement is adequate, regardless of individual words used
6. Where a statement's phrasing is ambiguous, the researcher must judge what meaning is contained (or inferred) within the statement and phrase it more clearly. This process should ideally be reviewed by other researchers to avoid bias as far as possible. If a statement is highly ambiguous it may be removed
7. If a statement does not refer [to the construct under investigation] it may be removed for inapplicability
8. It is the remit of the panel as a whole to decide if [concepts] in a statement are applicable. Therefore if a modification of a statement appears to bypass this process (i.e. by adding 'if applicable' to a statement) then, having considered the overall meaning of the statement, the researcher may remove this qualification
9. The overall goal of this process is to ensure simplicity and clarity and to avoid repetition, without any loss of the expressed concepts

4.2.2 Results

4.2.2.1 The panel

Following recruitment 76 people indicated a desire to participate. The breakdown of this into the sub panels can be seen in Table 4-2 along with the numbers who were contacted to see if they wished to participate, however it was not possible to accurately determine this for base trainers, as they were approached via another party, or for trainees as it was difficult to know the numbers present at the presentation.

Table 4-2. Recruitment and composition of sub-panels

Sub panel	Number approached	Number agreed to participate	Number completed round 1
'Expert' trainers	47	28	25
Base hospital trainers	-	14	14
Nurse endoscopists	36	19	19
Trainees	-	16	13
Totals		76	71

The number of participants who completed round one is also shown in Table 4-2; 93.4% of those who indicated a wish to participate completed round one. Looking at the panel as a whole, 32.4 % were female (23) and in terms of profession 26.8% were nurses (19), 52.1% physicians (37) and 21.1% surgeons (15); 66% (47) participants had attended a Training the Trainers course. The breakdown for each of the sub panels is shown in Table 4-3 and year of training for trainees in Table 4-4.

Table 4-3. Sub panel demographic

	Sex		Profession		Attended a 'training the trainers' course	
	Male	**Female**	**Physician**	**Surgeon**	**Yes**	**No**
'Expert' trainers	96% (24)	4% (1)	76% (19)	24% (6)	92% (23)	8% (2)
Base hospital trainers	92.9% (13)	7.1% (1)	71.4% (10)	28.6% (4)	85.7% (12)	14.3% (2)
Nurse Endoscopists	5.3% (1)	94.7% (18)			57.9% (11)	36.8% (7)
Trainees	76.9% (10)	23.1% (3)	61.5% (8)	38.5% (5)	7.7% (1)	92.3% (12)

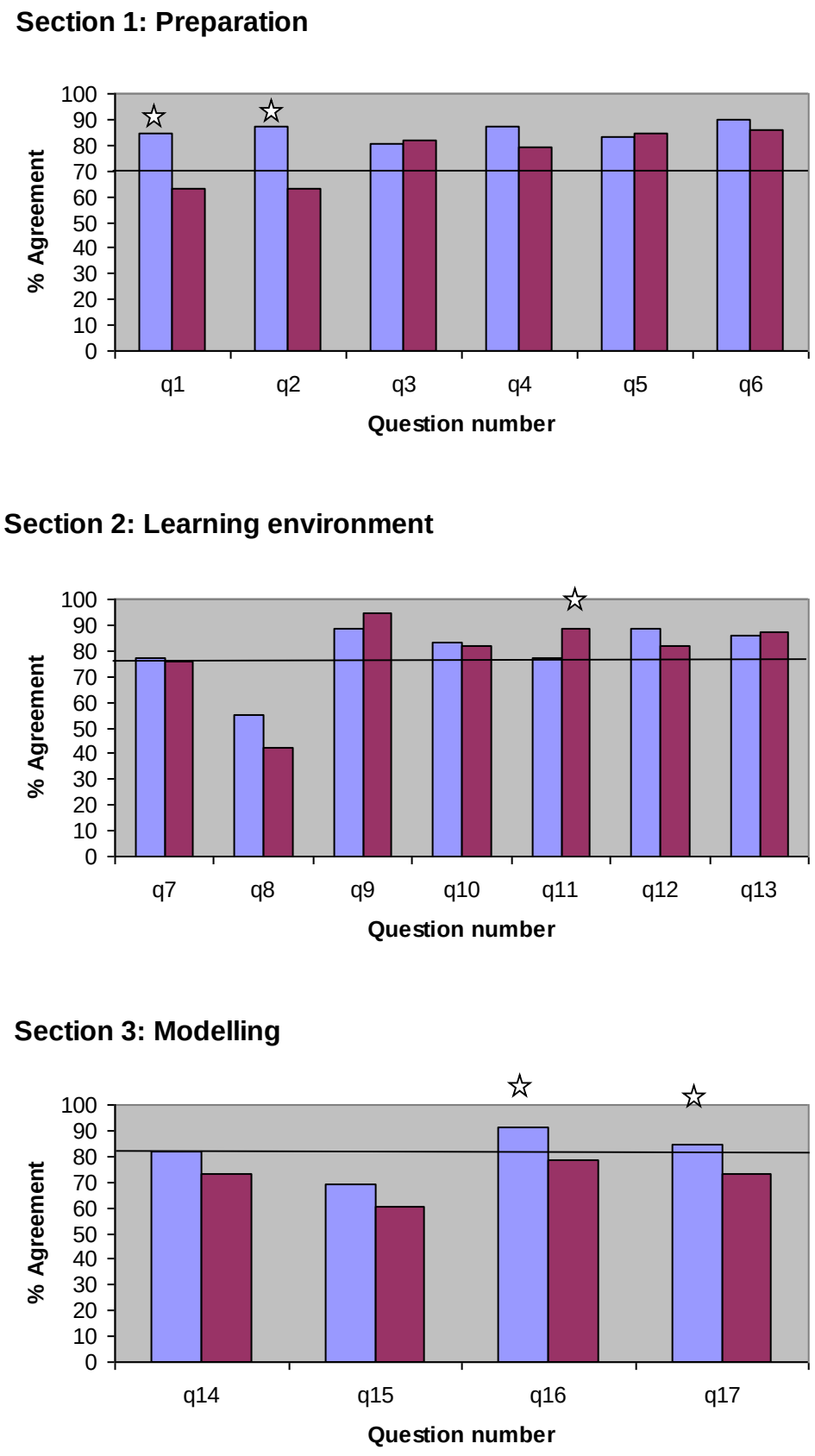
Table 4-4. Trainees by year of training

Year of training	**1**	**2**	**3**	**4**	**5**
Trainees	23.1% (3)	21.3% (3)	15.4% (2)	30.8% (4)	7.7% (1)

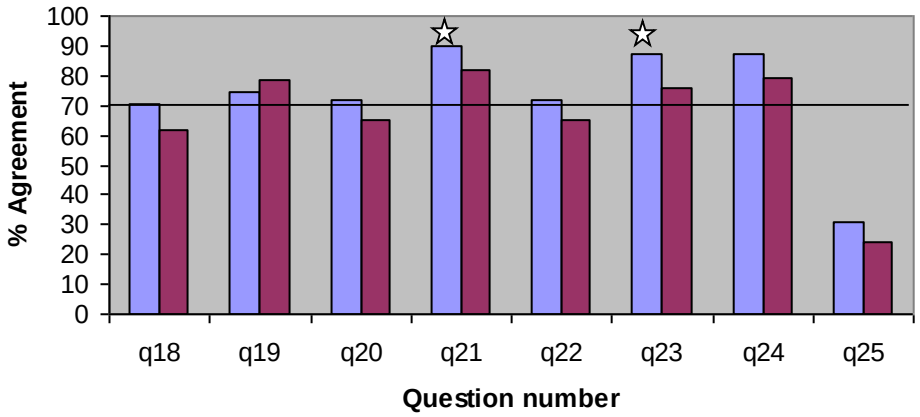
4.2.2.2 Quantitative analysis

Percentage agreements for the DOTS and LETS are shown in Figure 4-1. The attributes are separated into the categories in which Wells grouped them (discussed in chapter 3) as these were the sections in which they were presented to the panel. In Figure 4-1 the DOTS is in purple and the LETS in red. From this it is possible to see that 20 statements did not meet the criteria of 70% agreement required to be included in the tool kit. In those statements that did meet 70% consensus a star indicates a statistically significant difference between the score for the DOTS and the LETS meaning that the item has been allocated to that component of the toolkit without need for further review by the panel.

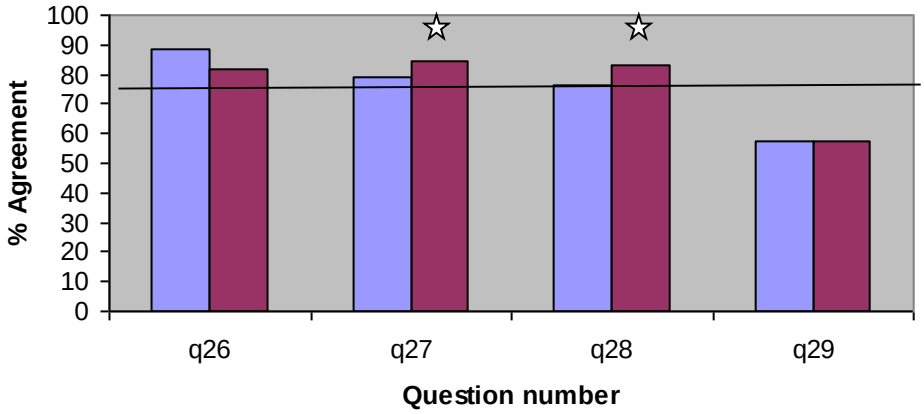
Figure 4-1. Percentage agreement for the items following round 1 of the Delphi.



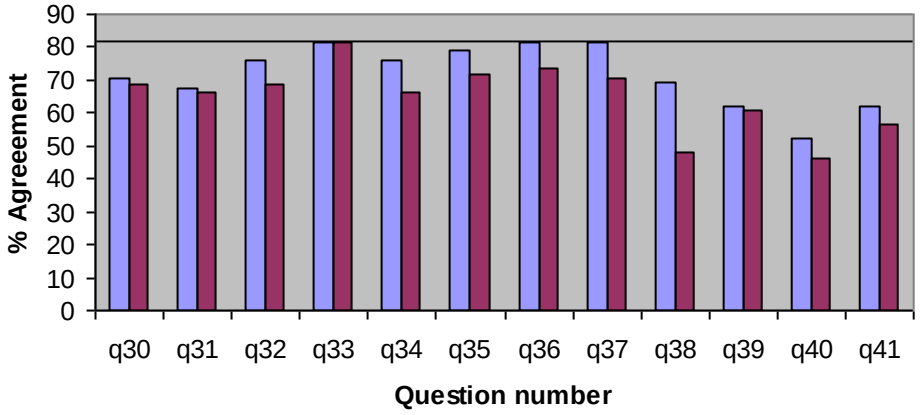
Section 4: Coaching



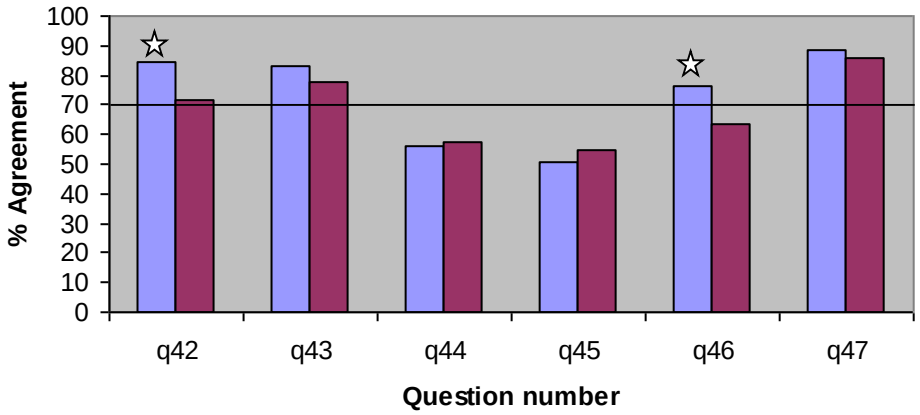
Section 5: Scaffolding



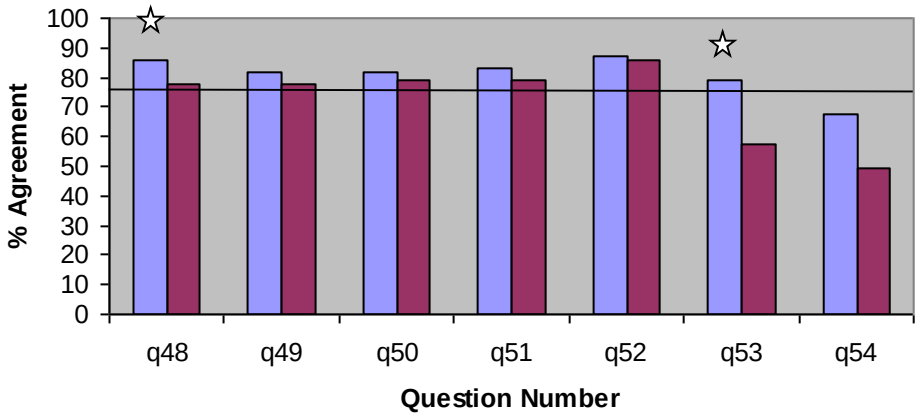
Section 6: Articulation



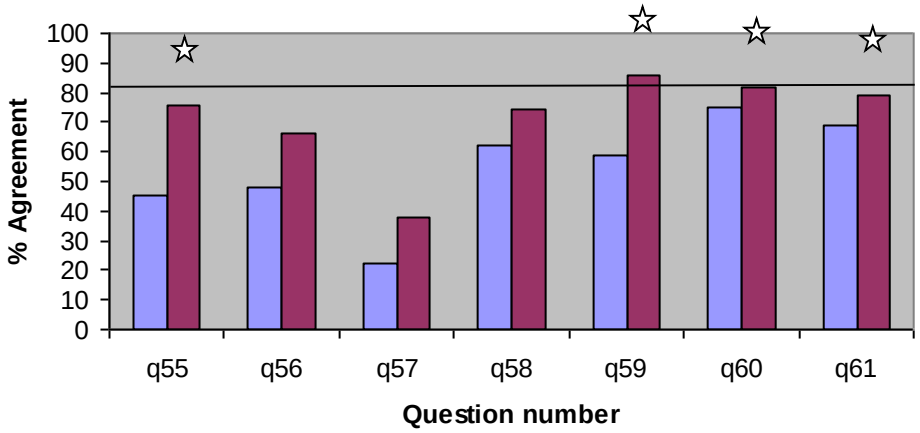
Section 7: Exploration



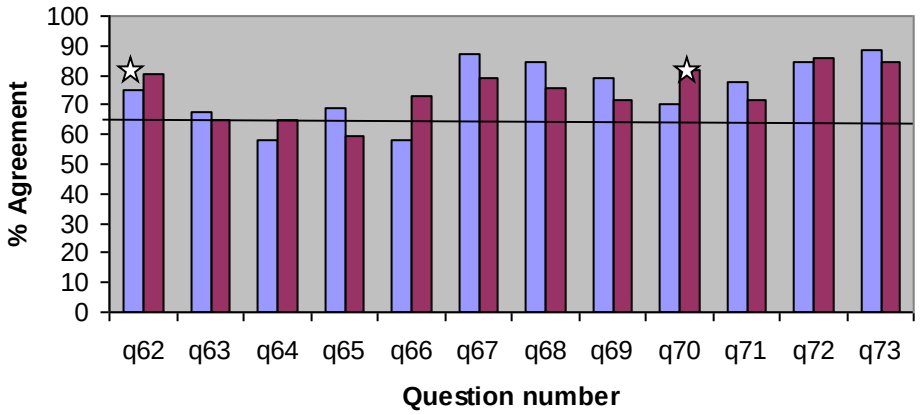
Section 8: Refection and Feedback



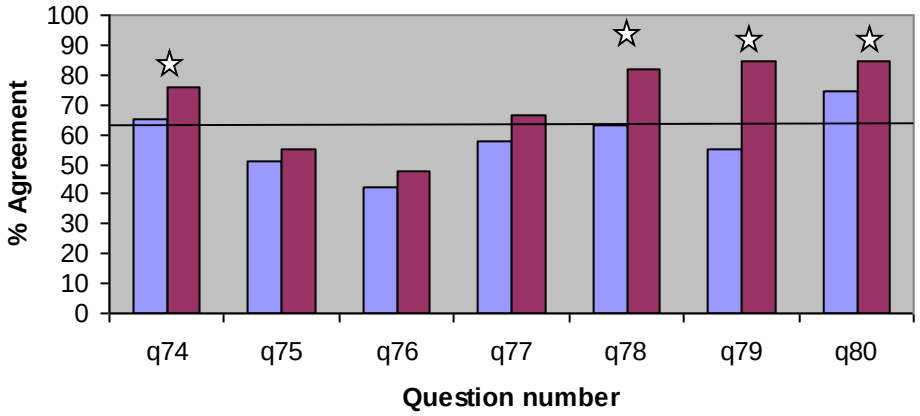
Section 9: Global



Section 10: Content



Section 11: Heuristic strategies



Key:

☆ Indicates difference between LETS and DOTS at p< 0.05 level of significance

DOTS

LETS

4.2.2.3 Qualitative analysis

Initially the free-text comments were reviewed by myself and SC for themes that occurred repeatedly across the sections and items. The themes that arose are listed in Table 4-5 and although these themes were not mentioned for every statement they were mentioned by several different participants for several different statements and therefore were taken as a set of generic criteria suggested by the panel by which all statements should be judged.

Table 4-5. Recurring themes across free-text comments

Recurring themes across the free-text comments
Too similar and required amalgamation
Statements should be grouped by theme more
Not measurable/rateable by the trainee
Not the trainer's responsibility
Statements required rewording
Some statements too over-arching and statements needed to be more specific
Statements were mutually exclusive

After considering the first two generic criteria that some of the statements were too similar and should be grouped by theme the statements were thematically analysed by LAM and SC and grouped into thirteen new categories, these are shown in Table 4-6. This was performed so that statements that were similar were grouped together to aid decisions about amalgamation and as a method of displaying the statements for the panel for round two. Each statement was then reviewed individually and as part of its section considering both the panel's generic criteria and the individual comments made for each statement. When considering any change to the wording of a statement the rules set out by Yeates (Yeates, Stewart et al. 2008) were used. LAM, SC and JRB were involved with the decision making and MRW then externally reviewed these decisions.

Table 4-6. New categories for items following round 1 of the Delphi process

New categories	
Rules and flow of session	Trainee's articulation
Goal setting	Trainer scaffolding
Intervention/observation	Interpersonal skills
Teaching strategies	Competence/Professionalism
Technical teaching	Team
Patient safety and comfort	Logistics
Feedback and reflection	

When the statements were re-categorised they were reviewed for similarity. As a result twelve statements were amalgamated into six; these are listed in the first section of Table 4-7 with the items amalgamated in the same row. In some places the wording of

the two items was combined to create a new statement, in others the wording of one of the items was retained and used if the concept of the second item was contained within the first item. If either of the items had met inclusion criteria for one of the components of the toolkit the new amalgamated item was retained in this component otherwise the items were included for review in round two. A further three items (items 10, 13 and 18 shown in Table 4-7) were excluded as although they did not measure entirely the same concept the concept was inferred and more measurable in another item.

During the review process three items that had not met the consensus threshold were re-included following review of their comments. For two of these items (item 63 and 40) this was because the panel, judging by their comments, felt the meaning was not clear. To ensure the correct meaning was conveyed these items were reviewed using the interview transcripts and Wells' (2010) N-VIVO nodes. The wording was then changed to try and more clearly reflect the intended interpretation; both the old and new wording is shown in Table 4-7. These items may have been rejected because they lacked clarity rather than the panel disagreeing with their content and therefore were included for review by the panel in round two. The third item that did not meet consensus which was included in round two referred to the trainer making physical contact with the trainee (item 25). Several of the panel commented that this may not always be appropriate and therefore the wording was changed to retain the original intent of the item about repositioning the trainee without the unacceptable element of inappropriate physical contact. This was included for review in round two.

In contrast some items that had met the 70% agreement threshold were subsequently excluded as they did not meet the set of generic criteria proposed by the panel. Two of these items were excluded as they were not possible for the trainee to assess (item 62, 66). Three items were excluded as these were not always within the trainer's control or their responsibility (items 3, 55, 56); this comprised the entire logistics category.

Finally as a result of reviewing the comments made by the panel the wording of some of the items was altered to add clarity. If these items had already been allocated to a component of the toolkit they did not require further review unless it was felt that the meaning may have been altered.

In reviewing the comments on the items that had been excluded by the research team prior to round one at least one participant thought that each item should be re-included. The only new attribute that was suggested and was felt to be assessable and generic to all stages of training and procedure was

'The trainer ensured the trainee produced accurate, comprehensive and easily understood reports'

Table 4-7. Items that have been altered or excluded as a result of analyzing the free-text comments in round one of the Delphi process

Items that have been amalgamated	
The trainer and trainee agreed and worked towards common objectives during the training period with a long term training plan (59)	The trainer and the trainee agreed goals for future sessions**.** (47)
The trainer adhered to the learning plan and reviewed the long term progress of the trainee (79)	The trainer knows the learning goals of the trainee and works toward these goals (80)
The trainer closely observed the process and was aware of what the trainee was doing' (24)	The trainer adjusted the position he was standing in the room appropriately, withdrawing as the trainee progressed (29)
The trainer identified aspects for the trainee to develop and improve.	The trainer delivered the feedback in a structure appropriate for the trainee (49)
The trainer clarifies everyone's role before a training encounter so that each individual knows how they are involved in the training process (1)	The nurses are informed it is a training list to ensure they are supportive of the trainee (2)
The trainer provided explanations at appropriate times (36)	The trainer was able to describe how he performed any endoscopic manoeuvres to his trainee that was understandable to the trainee and the trainee is left with an appreciation of how to perform the procedure (16)
Excluded items as concept covered elsewhere	
The trainer showed respect for the trainee (10)	
The trainer had realistic expectations for the trainee (13)	
The trainer taught the trainee to handle the scope gently (18)	
Consensus < 70% but items for review in round two with alternative wording	

Original wording	New wording
The trainer can explain the mechanics of endoscopy (63)	The trainer used their knowledge of the interaction between scope and anatomy to inform their training e.g. loop resolution
The trainer asked the trainee to demonstrate the problem when appropriate (40)	The trainer asked the trainee to show where they are struggling
The trainer physically moved the trainee to help them to achieve the desired outcome (25)	The trainer advised the trainee to move position to help them achieve the desired outcome
Excluded as not assessable by trainee	
The trainer taught according to the guidelines as per JAG and the DOPS (62)	
The trainer had a broad knowledge about the practice of endoscopy (66)	
Excluded as not always the responsibility of the trainer	
The trainer prepares the endoscopy list to meet the current needs of his trainee, both in volume and the nature of cases on the list (3)	
The trainer scheduled enough lists for the trainee (55)	
The trainer limited the number of trainees he was teaching to ensure that each trainee received adequate training (56)	

4.2.2.4 Discussion

The over-arching message from round one of this Delphi process was that the panel agreed with the majority of the statements; 75% of statements met the consensus threshold of 70% agreement. Perhaps this is not surprising if one considers the source of these statements; they were all derived from interviews from individuals belonging to the same four sub groups as the Delphi panel. The Delphi panel was therefore likely to have similar views. This first round adds validity to the work done by Wells (2010) as it adds breadth to the depth of his study.

Several of the comments left by participants noted the large number of items,

'very comprehensive – probably too detailed at present'

and one participant even repeatedly commented

'I agree that this is important, but I would to avoid the DOTS forms having so many domains as to make them very annoying to fill out regularly.'

One possible way of reducing the items further would have been to increase the level of agreement required to signify consensus. Doing this as a means to decrease the number of items may result in important concepts being lost (Stewart and O'Halloran 1999). In addition the aim was to create a formative tool therefore the detail is important in order to inform trainers as much as possible about their training(Harlen and James 1997).

An alternative method in order to reduce the items would have been to ask the panel to rank the items and this was even suggested by one panel member,

> *'It is difficult to disagree with many/most of these statements. Would ranking have been more appropriate?'*

This is an accepted Delphi technique (Okoli and Pawlowski 2004, Hsu and Sandford 2007). This would have had the advantage that it would have been possible to see which items the panel felt were most important, however ranking all 80 items would have been an onerous task and carried large participant burden (Streiner and Norman 2008) which is likely to have significantly reduced the response rate.

Throughout this process it was important that the final purpose and use of the toolkit was considered and that consistency was maintained with the criteria developed at the end of the cognitive interviewing process. These criteria were that the item must be measurable by all who complete the tool including a self-evaluation, a peer evaluation and a trainee evaluation, and that the item must be generic to any trainee regardless of the stage of training or procedure performed and for the DOTS that this must occur on every list. Several items were excluded prior to the Delphi process; when the panel were asked to review these items at the end of round one there was some support for each of these items. Despite this I have chosen not to re-include them as they do not meet the above defined criteria. The one exception to this is that loop resolution has been added as an example to the below item which was re-worded;

> *'The trainer used their knowledge of the interaction between the scope and the anatomy to inform their training e.g. loop resolution'*

4.2.2.4.1 Sub-group comparison

Participants for the Delphi process were recruited from the four different sub-groups. The data described above discusses the analysis of the data using the whole panel's

data as one group but it was also important to explore whether there had been any differences between the groups. As there were four groups it may have been possible to lose one group's viewpoint if it was masked by the other three groups especially as the groups were not of equal sizes.

This analysis by sub-group occurred after the second round of Delphi had been sent; this was because it was important not to let too much time lapse between rounds as I was concerned this may increase participant drop-out. The first questionnaire was available for six weeks therefore if someone had replied in week one there had already been a long lag even before any analysis started.

Initially a chi-squared test (or Fishers exact test for expected cell values of less than five) was performed for each item on SPSS to investigate for differences between groups. Using this technique identified any differences between how the subgroups had used the Likert scale for each item. On review many were differences between points next to each other on the scale for instance agree and strongly agree. These differences did not affect how I had handled the data within the Delphi and therefore were not of great interest. I was more interested in differences between the groups as to whether they agreed or disagreed with the items inclusion in one component of the toolkit. Data was therefore regrouped into two groups labelled one and two. Group one consisted of strongly disagree, disagree and neutral and group two comprised of agree and strongly agree. A chi-squared (or Fisher's exact) test was then performed again using these two groups.

This showed significant differences for four items (49, 63, 66, and 76). For those items that had already been included 70% of the whole group had already agreed or strongly agreed with the statement and this is enough to merit its inclusion even if one of the sub-groups disagreed with the item. In contrast for those items that had been excluded as a result of round one, one of the sub-group's believed that two items should be included; these were items 63 and 76,

'The trainer can explain the mechanics of endoscopy' (63)

'The trainer taught the theory of endoscopy before each new stage' (76)

Despite the fact that item 63 had only received 67.6% consensus for the DOTS and 64.8% for the LETS it had in fact already been included in round two, this was because

most of the criticisms had referred to its wording therefore, after reviewing the relevant transcripts from the original interviews, the wording was changed to

> *'The trainer used their knowledge of the interaction between the scope and the anatomy to inform their training e.g. loop resolution'*

and was included for review in round two. When I analysed the differences in this item 93.1% of the expert trainers felt it should be included. Its inclusion in round 2 is therefore also warranted on this level.

In regards to item 76, 17 of the 19 (89.5%) nurse endoscopists agreed or strongly agreed that this statement should be included in the LETS although it had been excluded by the group overall. I therefore decided on the basis of this that it should be included in the final toolkit, although the wording was changed slightly to reflect the comments from round one. The wording was changed to,

> *'The trainer checked the trainee's understanding of the theory of endoscopy before each new stage'*

One could argue that ideally this should have been submitted to round two to see if the whole group opinion changed but this was not possible as round two had already opened. Also I felt that the nursing opinion was so positively skewed that this should be acknowledged and the statement included. I feel it would be presumptuous to make comments about these group differences but one could speculate that this may be due to how the participants themselves had been trained either within endoscopy or their wider medical profession.

4.2.2.4.2 Missing data

When performing the analysis of sub-group data I realised a previous error had been made when round one was first analysed. Not all participants had reached the end of the questionnaire therefore there was some missing data. This had been noted during the data collection phase and those participants had been emailed to remind them to complete the questionnaire but not all had complied. As participants had reached different points in the questionnaire the number of missing data points varied by item. When the data had been analysed in SPSS in round one, a percentage for missing data had been calculated meaning that the other percentages were not of the participants

that had answered that question but rather of all that had started the questionnaire, which had not been intended.

All the percentages were therefore recalculated excluding missing data. As the amount of missing data varied from item to item this meant that the percentage agreement for each item increased but by varying amounts. Items that had been excluded because their percentage agreement fell below the threshold for consensus of 70% were then reviewed. Nine of the twenty items that had been excluded now reached above the 70% threshold for the DOTS, LETS or both. The highest percentage agreement excluding missing data for an item that had originally been excluded was 77.1% . It is important to note that significance values used to distinguish allocation between DOTS and LETS would not have been affected by missing data as it was performed using paired data only.

One of the options to overcome this problem was to raise the consensus threshold to 77.1% however as the amount of missing data varied from item to item some of the items that had been included had percentage agreements below 77.1%. Other options considered were removing these items as well or re-sending out the nine items that had levels of consensus above the 70% level taken as consensus. This latter option would have been difficult as round two was already underway and several participants had already completed it. The other argument against this methodology is that as I have already discussed the level of agreement was already very high and one of the aims of the Delphi process was to try and reduce the items to ensure that the toolkit is as useable as possible. Having discussed the issue with the research team I decided to consider raising the level deemed to signify consensus to 77.1%. In order to do this I needed to review those statements which would then also be excluded. This applied to three items; 18, 22 and 30.

'The trainer taught the trainee to handle the scope gently' (18)

had already been removed as it was felt that the construct was already measured elsewhere. The other two items were

'The trainer dealt with any slips, errors or mistakes made by the trainee' (22)

'The trainer agreed a common vocabulary with the trainee' (30)

And both had already been submitted for consideration in round two. A decision was made that these two items should be excluded regardless of the scores in round two. This means that consensus for round one was taken as 77.1%. This decision was also pragmatically advantageous as it meant that the workload of the participants was not increased. It does however show a limitation for the Delphi method that it is possible to increase or decrease the threshold which can alter a study's results.

4.3 Round 2

4.3.1 Methods

The data gathered in round one therefore informed the decisions regarding the management of round two. Given that the participants were clear that they wanted a short tool the aim of round two was to review those items that were deemed to have met consensus agreement but had not been clearly allocated to the DOTS or the LETS i.e. where there had been no statistically significant difference between the scores for the DOTS or the LETS. This was rather than also review all items that had not reached consensus criteria as well. Additionally participants commented that some of the attributes were too similar therefore the items were presented within the new categories identified (Table 4-6).

With this in mind the design of round two was considered; whilst the same design as round one could have been utilised this asked participants to consider the statement's suitability for the DOTS or the LETS separately whereas in this round the item had already reached the set criteria for inclusion and the question was to which part of the tool kit the statement should be included. Therefore in round two a similar question style to that used by Stewart and O'Halloran (1999) was used. Participants were asked whether they felt the item should be included in the DOTS, LETS, both or neither(Appendix 4). The reason for including a 'neither' option was that it enabled participants who had previously disagreed with the statement to have the opportunity to continue to disagree. A 'both' category was included as participants may feel that a statement could be included in either tool and I again felt it was important that they were able to select this as a preference.

Following the analysis of round one data thirty items were included in round two (although only twenty-eight statements were actually analysed as two items were subsequently removed following analysis excluding missing data). The questionnaire

was again delivered via surveymonkey (Surveymonkey 2008), a copy of this is in Appendix 4. The items were presented in their new categories from round one and at the top of each section participants were able to see which items from this section had already been allocated to the DOTS or LETS. This layout was chosen as it allowed participants to review items for similarity as this was frequently highlighted in round one. Participants were able to leave comments for each section rather than for every item.

One of the key features of a Delphi process, along with iteration, is feedback (Jones and Hunter 1995). Participants should be given the opportunity to review the statistics of the previous round in order to inform their opinion in this round; although Jones (ibid) warns that it should be made clear to participants that they do not need to conform to the group view. Participants therefore received a document summarising the results for round one along with the email informing them of the web address for the questionnaire. A copy of this summary is included in appendix 5. This document contained a brief description of how the results from round one had been analysed and a summary of the statistics and comments for each item. The statement summaries included the percentages of agree and strongly agree for the DOTS and the LETS; where the item had originated from, if the wording of the item had been changed, and a summary of the comments made by the participants from round one. Only a summary of the comments was given otherwise it was felt that the document would become too lengthy. The summary of comments was made by LAM and consisted of a combination of paraphrasing the comments and direct quotations. It was then reviewed individually by JRB, SC and MRW to ensure that the views in all comments had been adequately expressed in the summary and no bias introduced. Any discrepancies were then discussed and amended at a team meeting. Items that had been excluded were not included in the summary report as participants had no method of commenting on them and therefore were not included to ensure the report remained of a manageable size. In the instructions contained in both the email and on the questionnaire itself participants were encouraged to read the summary report but this was not enforced in any way.

Like Murry and Hammons (1995) only panellists who had completed the first questionnaire were sent the second questionnaire. The questionnaire could be completed at any time in the six week window with two reminder emails sent at two

week intervals. If the questionnaire had been started but not completed participants were also sent a reminder email to complete the questionnaire.

4.3.2 Results

Round two was open between the 24th May 2011 and the 11th July 2011. The 71 participants who completed round one were sent the questionnaire and 62 participants completed round two; this is an 89.8% completion rate for those who had previously participated. The composition of respondents by sub-panel is shown in Table 4-8 with the greatest attrition rate seen in the nurse endoscopist group.

Table 4-8. Composition of panel that completed round two of the Delphi process

Sub-group	Number completed round one	Number completed round two
Expert trainers	25	23
Base hospital trainers	14	14
Nurse Endoscopists	19	13
Trainees	13	12
Total	71	62

Table 4-9 shows the frequencies and percentage results for each domain by statement number.

Table 4-9. Results of round 2 of the Delphi process

Item	DOTS (%)	LETS (%)	Both (%)	Neither (%)
2B	24 (39.3)	6 (9.8)	31 (50.8)	0 (0)
2D	28 (45.9)	5 (8.2)	23 (37.7)	5 (8.2)
3C	40 (65.6)	2 (3.3)	17 (27.9)	2 (3.3)
4C	31 (50.8)	4 (6.6)	11 (18)	15 (24.6)
5B	11 (18)	20 (32.8)	28 (45.9)	2 (3.3)
5C	32 (52.5)	1 (1.6)	20 (32.8)	8 (13.1)
6A	17 (27.9)	8 (13.1)	31 (50.8)	5 (8.2)
6B	14 (23)	6 (9.8)	39 (63.9)	2 (3.3)
7C	23 (37.7)	4 (6.6)	33 (54.1)	1 (1.6)

7D	22 (36.1)	4 (6.6)	31 (50.8)	4 (6.6)
7E	27 (44.3)	4 (6.6)	27 (44.3)	3 (4.9)
7F	14 (23)	5 (8.2)	38 (62.3)	4 (6.6)
7G	7 (11.5)	16 (26.2)	37 (60.7)	1 (1.6)
8C	14 (23)	8 (13.1)	37 (60.7)	2 (3.3)
8D	11 (18)	12 (19.7)	37 (60.7)	1 (1.6)
8E	7 (11.5)	12 (19.7)	41 (67.2)	1 (1.6)
10B	12 (19.7)	6 (9.8)	38 (62.3)	5 (8.2)
11C	13 (21.3)	15 (24.6)	29 (47.5)	4 (6.6)
11D	3 (4.9)	30 (49.2)	26 (42.6)	2 (3.3)
11E	9 (14.8)	18 (29.5)	28 (45.9)	6 (9.8)
11F	7 (11.5)	12 (19.7)	38 (62.3)	4 (6.6)
12A	12 (19.7)	11 (18)	34 (55.7)	4 (6.6)
12B	22 (36.1)	8 (13.1)	26 (42.6)	5 (8.2)
12C	29 (47.5)	6 (9.8)	23 (37.7)	3 (4.9)
12D	29 (47.5)	6 (9.8)	20 (32.8)	6 (9.8)
12E	32 (52.5)	5 (8.2)	19 (31.1)	5 (8.2)
13D	24 (39.3)	6 (9.8)	24 (39.3)	7 (11.5)
13E	21 (34.4)	6 (9.8)	26 (42.6)	8 (13.1)

4.3.2.1 Managing the statements

When looking at the results from round two, no clear way to analyse the data was apparent immediately. No item received more than 70% agreement for its inclusion in one tool or both which had been deemed as consensus in round one. However the use of the 'both' option confounded this issue as those respondents that ticked 'both' could be argued as having indicated that they wanted it in the DOTS and the LETS and therefore these scores could be added to those who clearly stated that they felt a statement was suited to one element of the toolkit only. Table 4-9 shows the frequency results for each statement; 19 of the statements scored more in the 'both' category than for either the DOTS or LETS. Although the 'both' category had been included in order to allow participants to demonstrate that they felt the statement was equally suitable for both elements of the toolkit if all these items were included in both tools then each tool would become unwieldy. Those completing it may also become frustrated by answering the same questions on the two different tools and trainers would receive the same feedback from both tools which would confer no advantage.

Therefore a decision was made not to include any items in both tools, rather make a decision as to whether it should be included in the DOTS or LETS

Having made this decision the 'neither' category was reviewed next. As can be seen from Table 4-9 only one statement (2B) achieved 0% in the neither category. This means that for all the rest of the items at least one participant felt that it should not be included in the final toolkit. For roughly a third of the statements this was just one or two participants however for four statements over ten percent of the panel felt that it should not be included. The comments about these statements were then reviewed. Nearly a quarter of the participants felt that the statement

> *'The trainer's attention to each moment of the procedure was appropriate to the trainee's needs' (Q4C)*

should not be included in the toolkit. The comments in reference to this statement referred to feelings that this was overly intense and this appeared to be why this statement was rejected. One participant did suggest changing the words to the 'conduct of' but this would change the meaning of the statement outwith that intended.

> *'The trainer advised the trainee to move the position to help them achieve the desired outcome' (Q5C)*

This statement in fact fell significantly below the 70% threshold in round one and had been substantially reworded after round one to overcome participants concerns; participants in round two felt that the wording was now too vague.

The other two statements that received greater than ten percent in the neither category were

> *'The trainer reassured the trainee at appropriate times'*

And

> *'The trainer helped the trainee to carry out the procedure independently'*

Both received comments about being vague and the latter was also criticised for being ambiguous. From the comments about all of these statements there were no specific suggestions for change that would overcome other participants' criticisms. A decision

was therefore made to take ten percent of the scores in the neither category as a cut off and those that received greater than ten percent were to be excluded. The above discussion has shown how the rules to manage the decision making process for round two developed; these rules are summarised in Table 4-10.

Rules for managing statements in round two.
Omit all statements that >10% of participants had marked as 'neither'
No statement to be included in both tools; therefore omit the 'both' category in analysis
Allocate statements to the DOTS or the LETS depending on the highest scores

Table 4-10. Rules for managing statements in round 2.

Items were then allocated to either the DOTS or LETS based on which category they had scored highest. Relative risk was calculated to review the difference between the DOTS and LETS. This was calculated by dividing the DOTS by the LETS; therefore large numbers suggested a strong preference for the DOTS whereas very small numbers suggest a strong preference for the LETS. Three items' relative risk ratios were very close to one i.e. no strong preference for the LETS or the DOTS; these were items 8D, 11C and 12A. These items were therefore reviewed separately.

'The trainer identified aspects for the trainee to develop and improve' (8D)

This item received one more vote for the LETS than the DOTS; however this item refers to the process of feedback and fits very closely with the preceding statement,

'The trainer reinforced positive aspects for the trainee to develop and improve' (8C)

which was clearly allocated to the DOTS. Additionally educational theory suggests that feedback should occur close to the event (Brinko 1993, Richardson 2004) and regularly (Bing-You and Stratos 1995) in order for the trainee to continuously improve. Based on this theoretical basis I opted to include this item in the DOTS.

The other two items which the panel appeared divided as to whether they should be include in the DOTS or the LETS were items 11C and 12A. In reviewing the comments made in both rounds for these items there was no clear reason why these items should be allocated to either tool. Rather than make a decision at this time I opted for these

two items to be allocated into both components of the toolkit and will examine these further in the next chapter when the list of items are converted into actual tools.

Lastly the comments were reviewed to explore further thoughts about the items. There were markedly fewer comments in this round than round one, and comments that were made largely reflected those that had already been made in round one. One of the comments made in the teaching strategies section was

> *'If you have dozens of statements of things that are obviously necessary it just makes these assessment tools very off putting and excessively time-consuming to complete'*

This section did list a lot of skills that a trainee endoscopist needs to acquire and therefore are important to teach but it may not be necessary to concentrate on each of them in every session. One could therefore argue that despite the fact that all these items received a much higher percentage in the DOTS this could be overcome by placing the statements in the LETS. However the skills that the trainer will need to teach will also depend on the prior experience and level of the skill of the trainee. It was therefore decided to keep this concept within the DOTS but combine the statements into one overarching statement with specific examples,

> *'The trainer gave specific skills teaching (e.g. keeping luminal view, examine the mucosa, tip control, appropriate insufflation, loop resolution)'*

No other changes were made to the statements based on the comments made in this round.

4.4 Discussion

At the end of round two all but two statements were either allocated to an element of the toolkit or excluded. This negated the need to conduct a third round of the Delphi process as it would have been unlikely to lead to further consensus. It may have been possible to conduct a third round of the Delphi process for those items that were still not clearly allocated to one component of the toolkit but it was felt there was a danger of being too driven by statistics to allocate the items which may mean that the tools do not actually make sense. This has been illustrated above when discussing the feedback items, if the decision making process had been driven by statistics alone then item 8C and 8D which theoretically go together would have been allocated to different

components of the tool. A third round could have also been conducted without the 'both' category to try and gain further consensus about the items however those that felt an attribute was equally suitable to both tools did agree that the item should be included in the toolkit. Whilst I previously stated that the literature suggests that three rounds is optimal (Villiers, Villiers et al. 2005, Hsu and Sandford 2007) as I had already pre-determined the list of attributes then this reduces the number of rounds required (Stewart and O'Halloran 1999).

The use of a 'both' option within round 2 became difficult to manage, I had included it as I felt that it was feasible and permissible for members of the panel to believe that an item was equally suited to both tools. However the 'Both' option was often used preferentially rather than participants selecting a preference for an element of the toolkit. This could be for several reasons; one reason may be that participants genuinely felt that an item should be included in both tools; or were unable to decide which element it should be included in. An alternative explanation was that it may have been easier to choose 'both' than truly consider the best placement for the item. Although for some items it may have been very appropriate to include it in both elements, repeatedly participants in both rounds emphasised the need for the toolkit to be as short as possible. It is likely that those with this view would also feel that people would be put off using a toolkit that contained the same questions in both elements and I felt it was important to acknowledge and act on this viewpoint. Additionally in trialling the toolkits I aim to see if the two tools correlate; inclusion of the same questions would inflate the correlation.

This also highlights one of the limitations to a Delphi process that it relies on one's group of experts to fully consider and reflect on each statement before offering their opinion and that they fully understand the aim of the study. One way I tried to achieve this is to get participants to indicate their interest in the study before sending them the first questionnaire. The implication was that only those interested in the study would agree to participate and were fully engaged in the subject matter. However, an expression of interest may not indicate attention to detail or a clear understanding of the instructions.

As discussed at the beginning of this chapter choice of expert is important as it has the potential to alter the outcome of the process. In choosing the groups of experts I wanted to capture the opinions of those that would be later using the tool, the reasons

for this is that I felt they would have the most valid opinions on what attributes were most important in capturing the high-quality endoscopy trainer. I also felt that using the groups that would later use the tool might increase 'buy in' (Moore 1987 cited in Clayton 1997) and the eventual adoption of the toolkit. Other groups however may also have had useful inputs into the Delphi. One of the groups that I could have used were patients. According to Wells' (2010) model all teaching should be patient-centred therefore one could argue that patients are ideally placed to ensure this remains the case. One of the issues with this though is that if one were to include patients within the Delphi they would have needed to experience an endoscopy procedure. Patients are often lightly-sedated during procedures and therefore their recall of the event maybe impaired particularly to subtleties in teaching practices. Not all teaching such as feedback might happen in the patient's presence. I also felt that in terms of some of the teaching skills such as scaffolding and fading a patient may not be able to identify and appreciate the importance of these during limited endoscopic exposure. I therefore feel that although patients may have rated items that ensured that teaching remained patient centred highly there may have been areas that they were not able to evaluate so easily and this may have skewed the Delphi process. The other group that could have been included in the Delphi process was those nurses that work in endoscopy as assistants. Unlike patients they will have had exposure to multiple teaching episodes in endoscopy often by many different trainers with different trainees. As they are not involved in the teaching process directly they may have valuable observations as an observer. They are likely to have witnessed both effective and non-effective teaching moments. In addition their priority within endoscopy remains with the patient and they are likely to be able to identify those behaviours that ensure that teaching remains patient-centred. Endoscopy nurses may not always witness all of a teaching episode but this does not mean that they would have not recognised the importance of feedback. An argument for not including endoscopy nurses is that because they have not learnt or are in the process of learning endoscopy then they may not appreciate the importance of some training techniques. I opted not to include endoscopy nurses or patients as groups in the Delphi as they were not included in Wells' original research; I hoped that by using the Delphi process I wished to add breadth to the depth of his work, I felt that by using different groups this would not have added as much strength to his work.

Alternatives to the Delphi process were discussed at the beginning of the chapter and despite the difficulties that arose as part of the Delphi process I still felt that it was appropriate group technique to use. I have discussed that rather than rating the items I could have asked the panel to rank the items however given the large number of items initially this may have led to cognitive overload and would need to have been repeated, once for the DOTS and once for the LETS. I therefore wondered whether using an alternative rating scale may have been preferable. A tool designed to give feedback to general practice trainers (Donner-Banzhoff, Merle et al. 2003) used the importance-quality score method. This is a method derived from health-related quality of life research in which an importance-severity score is utilised to determine which items are to be included on a tool to ensure that those measures that affect patients the most and are seen as most important are included. In a similar manner Donner-Banzhoff, Merle et al (2003) asked General Practice trainees to rate the quality of their trainer and the importance of that item to their training over 121 different items. To create an importance-quality score the inverse of the quality score was added to the importance score. The lowest scoring items were retained. In this way items that were rated most important by the trainee and of lowest quality (i.e. the items where there was room for improvement) were most likely to be retained. I felt that this could have been an alternative scoring system to that used in the Delphi and may have given the panel more to reflect on when reviewing the items; however there are also disadvantages to this method. This method means that only the lower scoring items in terms of how a trainer performed are retained. Whilst the rationale for this is that it reduces items that are always scored highly, it could lead to the trainer perceiving that they are worse trainers than they actually are and that important attributes are excluded. The authors acknowledged this latter point and in fact kept some items on their 'importance' score alone.

In the Delphi process each statement was reviewed individually by the panel; this does mean that each statement was accepted or rejected on its own merit. In round two the attributes were presented within the new categories with the attributes that had already been included in the toolkit listed at the top of the section. This may have altered participants' judgement but it is not possible to know if this was acknowledged by the panel. Certainly there was no explanation given to how the attributes fit together or a description of Well's model of effective endoscopy teaching. This may

have affected the resulting list of attributes as participants may not fully appreciate all the attributes.

In order to try and influence the results of the Delphi as little as possible, particularly in terms of the qualitative analysis of the free text comments, I tried to utilise 'rules' in order to make decisions about each of the items in a consistent manner. In terms of making decisions about how to alter the wording I used the rules set out by Yeates et al (2008) as listed in Table 4-1. This list of rules was helpful as they had also dealt with a large number of items that needed to be reduced and amalgamated and it was helpful to have rules in order to govern these items. As well as using the statistics to determine an item's inclusion or exclusion in the toolkit I also used the comments that were repeatedly made by the panel in order to form a list of generic criteria. This was to try and create some consistency in opinion and how decisions were managed. All decisions about items were made by myself with at least one supervisor and were then reviewed by another supervisor in order to try and ensure objectivity and consistency was maintained. By creating and using a set of rules to govern these decisions this would mean that if another person were to analyse the results then they would be able to follow the process in a similar manner and produce similar results. It is worth noting that although I used the panel's comments to form generic criteria the panel did not necessarily use these criteria consistently themselves. For instance they did not necessarily make these comments for every item that I then applied them to. Additionally the number of comments made varied from panel member to panel member and by using these comments to create a list of generic criteria these might in fact represent the views of a few dominant individuals (or dominant on paper in any case). They however do allow for consistency and a degree of objectivity.

As previously mentioned this Delphi process has added to the validity of the work done by Wells (2010) due to the high levels of agreement with the attributes by the panel. It has also contributed to the content validity of the toolkit by involving a broader constituency of stakeholders who have important views on training. As a result of this process the wording of some of the attributes has changed and some of the items have been amalgamated. This may have changed the meaning of the items slightly from that intended by Wells (ibid). This does not necessarily decrease the content validity of the toolkit as the argument is that, as this has been developed using stakeholder opinions, the attributes have been improved. It is important that this has occurred in a

transparent and logical manner hence the use of the above results to ensure that the process is clear.

4.5 Conclusion

The final DOTS and LETS statements are shown Table 4-11and Table 4-12 along with the round in which they were allocated. In the next chapter I will discuss how the tools themselves were created.

Table 4-11 Statements allocated to the DOTS at the end of the Delphi process

DOTS		
Statement	**Round in which allocated**	**Amalgamated statement or wording altered**
The trainer agreed objectives for the session	Round 2	Wording altered
The trainer ensured the trainee knew the name and role of each member of the endoscopy team before a teaching encounter	Round 1	Amalgamated statement
The trainer agreed and applied the ground rules including when to intervene	Round 2	Wording altered
The trainer questioned the trainee at appropriate times	Round 2	
The trainer provided explanations and descriptions at appropriate times	Round 1	Amalgamated statement
The trainer used a mixture of suggestions, prompts, solutions and instructions	Round 1	Wording altered
The trainer checks the trainee has understood what has been said through observation and direct questioning	Round 2	Wording altered
The quantity of dialogue was appropriate for the trainee and specific teaching episode	Round 2	
The trainer actively listened to the trainee	Round 2	Wording altered
The trainer asked the trainee to show where they are struggling	Round 2	Wording altered
The trainer gave specific skills teaching (e.g. keeping luminal view, examine the mucosa, tip control, appropriate insufflation, loop resolution)	Round 1 and 2	Amalgamated statement

The trainer did not overburden the trainee with too many tasks	Round 2	Wording altered
The trainer demonstrated a procedure where necessary	Round 1	
The trainer intervened in a timely fashion (either at a predefined time or if the trainee was struggling)	Round 1	
The trainer allowed the trainee reasonable time to carry out the procedure	Round 1	Wording altered
The trainer always ensured that the patient was comfortable and safe and their dignity was maintained	Round 2	
The trainer encouraged the trainee to communicate appropriately with the patient	Round 2	
The trainer helped the trainee to assess if the objectives for the session had been achieved	Round 1	
The trainer reinforced positive aspects of the trainee's performance	Round 2	
The trainer identified aspects for the trainee to develop and improve	Round 2	

Table 4-12 Statements allocated to the LETS at the end of the Delphi process

LETS		
Statement	**Round in which allocated**	**Amalgamated statement or wording altered**
The trainer made the trainee feel welcome	Round 2	
The trainer agreed and worked towards common objectives during the training period with a long term training plan	Round 1	Amalgamated statement
The trainer matched their approach and pace to the trainee's needs (needs defined by stage, preferred learning style, level of confidence)	Round 1	Wording altered
The trainer used teaching aids that can support learning (e.g. the magnetic imager, diagrams, models etc.)	Round 2	

The trainer took advantage of opportune moments to teach	Round 1	
The trainer checked the trainee's understanding of the theory of endoscopy before each new stage	Round 1	Wording altered
The trainer taught the whole process of endoscopy e.g. the indications, consent, communication and sedation	Round 1	Wording altered
The trainer ensured accurate, comprehensive and easily understood reports were produced	Round 2	New statement suggested by participants
The trainer actively listened to the trainee	Round 2	Wording altered
The trainer was patient and calm	Round 2	
The trainer was available and focused on the trainee – by minimising distractions	Round 1	Wording altered
The trainer developed a good working relationship with the trainee	Round 2	
The trainer set a good professional example through their own behaviour	Round 1	
The trainer built the trainee's self confidence	Round 1	Wording altered
The trainer reviewed the data collected by the trainee to inform feedback e.g. DOTS forms, CuSum etc	Round 1	Wording altered
The trainer helped the trainee to reflect on the trainer's performance	Round 2	
The trainer reviewed the trainee's long term progress	Round 1	Amalgamated statement

Chapter 5. The Final Toolkit

Following the Delphi process items had been allocated to the LETS and DOTS components of the toolkit; the list of items now required transforming into a toolkit. By this I mean that following the Delphi process I had a list of items to be used in the two tools but these were only lists. In order to create a toolkit these items needed response options and instructions as to how to be completed. In this chapter I consider the various options that could have been used to form response options and decide upon the use of a Likert scale. The wording of some of the items had also changed significantly during the Delphi process and I wanted to ensure that the items still had clarity; I therefore trialled the tool with trainees and a trainer and performed a further round of cognitive interviewing to examine this further.

5.1 Designing the toolkit

As mentioned above one of the most important considerations following the Delphi process was to convert the two lists of items into evaluation tools. One of the most important considerations for this was to consider how the items should be scored. As the aim was to create a toolkit that could be used by peers, trainers and as a self-evaluation tool for trainers I wanted the response options to be the same for all three groups. One of the reasons for this is that the evidence for the use of self-evaluation is that discrepancies between how a trainer scores themselves and a trainee scores them can act as a powerful motivator for change (Stalmeijer, Dolmans et al. 2010) I therefore wanted to enable a direct comparison between scores so it was important that the response options were identical.

To convert the lists of items into a usable tool it was important to consider those areas that might influence the response process category of validity evidence (American Educational Research Association, American Psychological Association et al. 1999). Clearly if the instructions to a tool are not clear or the response options are not appropriate this will influence how the tool is completed and is likely to detract from the construct under investigation. As previously discussed descriptions of the development of other evaluation tools provide very little evidence for this category of validity evidence. This may be because previous studies gave very little consideration to these factors or because it is difficult to provide concrete evidence for this area. Ideally

I would have created the toolkit using several different types of response options then go on to trial these but this would have been difficult within the available timescale. I therefore opted to use the literature and make a decision on the scale on this basis.

5.1.1 Scaling the items

In deciding how the items should be scaled it was important to consider the nature of the items; all of the items ask the respondents to make a decision about whether the trainer has, or displays, that ability or skill; this decision is in fact a continuous variable where each respondent can agree or disagree to any extent; however this agreement needs to be converted to a scale. One option could be just to ask the trainer to say yes or no to this question i.e. they agree or disagree. An advantage of this dichotomy is that it forces the respondent to make a choice and either agree or disagree with the statement, because only this choice needs to be made it has been argued that dichotomous scales are quicker to complete (Clark and Watson 1995) however it also has disadvantages. As the respondent only has two options it leads to a loss of information(Streiner and Norman 2008), as it would not be possible tell if the respondent strongly agreed with the statement or just mildly agreed. Also if the results are very skewed in one direction (i.e. everyone either agrees or disagrees with an item) it can lead to distorted correlational results (Clark and Watson 1995). Although, it is important to acknowledge that the variable being measured is continuous, there still needs to be a method by which these judgments are quantified; one way of doing this is to use direct estimation methods.

5.1.1.1 Types of scale

Direct estimation methods are designed to elicit from the respondent 'a direct quantitative estimate of the magnitude of a variable' ((Streiner and Norman 2008) p41). There are several methods by which these direct estimations can occur.

5.1.1.1.1 Visual Analogue scales

A visual analogue scale is where a respondent is asked to make a mark on a line of fixed length (normally 10cm) with descriptive anchors at each end; for instance in a scale about pain the anchors used might be 'no pain' and 'worst pain ever' or in our case on the most global setting 'best trainer ever' to worst trainer ever'. One argument for this method is that it can allow for greater precision as the mark can then be measured to two decimal places (if measured in millimetres) and therefore have an apparent

accuracy of one percent; however there is no guarantee that those respondents who are using that scale are making their mark with the same accuracy. There has been lots of work looking at the use of visual analogue scales to measure perceptions of pain and quality of life but less so in the field of medical education (McLean 2001).

5.1.1.1.2 Likert scales

Rather than asking a respondent to mark their opinion along a continuum a Likert scale (Likert 1952) gives a respondent a choice of prescribed options to select. These options are along a bipolar range. For instance, Likert scales are often used to measure agreement and the scale may range from strongly disagree to strongly agree but can measure any attribute such as frequency or acceptance. Likert scales are very common within clinical evaluation tools; all of the tools that were evaluated by Fluit el al (2010) utilised a Likert scale. Streiner and Norman (2008) note two important points about Likert scales; one is that the adjectives used should always be appropriate for the stem (i.e. the statement) and secondly it is important to ensure that the mid-point is labelled in such a way that it reflects the midpoint of the attribute rather than just an inability to answer the question.

The goal of the above techniques is to assign a numerical value to an item. Although the scale always increases in value, regardless of whether it is unipolar or bipolar, the data produced is not interval data, rather it is ordinal data. The reason for this is that the distance between one interval and the next may not be identical, for instance the difference in neutral to agree may not be the same as the difference in feeling between agree and strongly agree(Jamieson 2004). This is important as it affects the statistical tests that can be used, as it is technically incorrect to perform statistical tests that are for interval data (parametric tests) on data derived from ordinal rating scales; however it is generally felt that as long as the data is not significantly skewed then this is acceptable (Streiner and Norman 2008).

Celana and Roberts (2011) conducted a study which compared the use of Likert scales to visual analogue scales to evaluate a teaching programme within an emergency department. The authors hypothesise that because visual analogue results are used regularly within the emergency department in which the study was conducted, participants should be familiar with them. The authors used the same questions to measure perceptions of different teaching programs within the emergency department (Celana and Roberts 2011) but asked participants to answer the questions both on a

Likert scale and using a visual analogue scale. The authors measured the test-retest reliability of the tools for both scales using the intraclass co-efficient. The results suggested a slightly higher test-retest reliability for the Likert scale than the visual analogue scale but this is not tested for statistical significance. The correlation between the visual analogue scale and the Likert scale for the individual questions was also calculated; the two scores had a significant correlation for 17 of the 26 questions. This suggests that the two rating methods appear to give similar answers but does demonstrate that this was not the case for all questions. The authors do not explore which rating method is superior or gives a more accurate representation when the correlations differed. This highlights that using different rating methods may lead to different scores. The authors comment that inputting the data for the questionnaire that used the VAS took longer due to having to take measurements for each question.

5.1.1.1.3 Alternatives to rating scales

An alternative to rating scales altogether would be to ask the respondents to rank the items. Ranking items forces respondents to differentiate between items however it is more cognitively difficult to ask participants to rank rather than rate (Streiner and Norman 2008). Ranking would not be suitable for the evaluation tool as all the attributes on the evaluation tool are desirable and therefore although it may be useful for a trainer to know in which areas they excel (i.e. ranked most highly) it would be difficult to interpret the meaning of lower ranked items in terms of whether they just excel at the higher ranked items or really are bad at the lower ranked ones. Also when feeding the results back to trainers aggregated scores may not indicate extreme values where trainees had assigned a different order for the attributes and so would provide less information. Ranking the items would also be a more time consuming process as the respondent would have to review every item in relation to every other item rather than considering each of them individually. This may confer an advantage in forcing the respondent to consider the items more closely,and might be particularly true for the items ranked highest and lowest.

5.1.1.2 Scale choice

I opted to use a Likert scale for both elements of the toolkit; this was because results between Likert scales and visual analogue scales appear to be similar when considered in the field of medical education (McLean 2001) but the results of Likert scales are easier to calculate and can be inputted more quickly. Anchor points on Likert scales

also ensure that the meaning described to a point is roughly the same between raters. Additionally Likert scales are common amongst teacher evaluation tools (Beckman, Cook et al. 2005, Fluit, Bolhuis et al. 2010). Endoscopy trainees are used to using Likert scales both in terms of the previous endoscopy trainer evaluation tool but also that the Directly Observed Procedural Skills (DOPS) form used to assess trainees also uses a Likert scale(JAG 2012).

Having decided to use a Likert scale I next had to decide how many scale categories (i.e. points along the scale) to use. If the number of categories is less than the rater's ability to discriminate then this will result in a loss of information; however if there are too many intervals then this is likely to make the tool more cumbersome and time-consuming. The reliability of the tool decreases as fewer categories are used(Streiner and Norman 2008). The loss in reliability when category numbers are reduced from ten to seven is small; however, if the number of categories is reduced to five the reliability decreases by 12%; if only two categories are used to rate a continuum the reliability decreases by 35% (Nishisato and Torii 1970) cited in (Streiner and Norman 2008). This however has to be balanced with respondent preferences and ease of use. Respondents appear to dislike it if they are given too few categories on the scale (Streiner and Norman 2008) and were found to undergo cognitive overload if there were too many; therefore the optimal number of categories appears to be between five and seven. Additionally it is argued that increasing the number of scale categories may in fact reduce validity if the respondents are unable to make the more subtle distinctions that are required (Clark and Watson 1995). I opted to use five categories, as I wanted to ensure that I captured enough detail whilst ensuring the tool was easy to use. Also because reliability is also a function of the number of items on a tool (Field 2009) and each tool is comprised of several items reducing the number of categories should not unduly effect the overall reliability (Streiner and Norman 2008).

I opted for an odd number of categories; a Likert scale does not necessarily need to have an odd number or have midpoint suggestive of neutrality; removing a neutral point can mean that the respondents are forced to make a decision. Whilst I was concerned that a middle neutral category can be ambiguous I felt that teaching did not necessarily need to be either good or bad and that therefore it was reasonable to leave a neutral midpoint. I also chose to have this neutral midpoint in the middle; as in teaching evaluation students sometimes score only using the upper half of the scale

('the halo effect') then the bottom half of the scale becomes redundant, however in this initial trial phase I had no evidence that this would occur in the evaluation of endoscopy trainers and therefore opted to use a balanced scale.

The next stage in scale development was to decide what descriptors should be added to the scale; this was essentially asking what it was about the attributes that I wanted respondents to make a judgment on. For ease of use I wanted all the items to use the same adjective in each tool but both tools did not necessarily need to have the same adjective. This decision was made in order to improve the tool's utility but meant that a compromise had to be made when choosing an adjective that was appropriate for all items rather than the most descriptive adjective for each item. The two options that I felt would be potentially suitable were frequency or agreement. In order to make a decision I reviewed the items. In the LETS because there were more items of an interpersonal relationship, such as developing a working relationship, I did not feel it was appropriate to ask how often the trainer did these and that the degree to which they did it was much more suitable, therefore degrees of agreement was chosen as the adjective. This also reflects the fact that recall decreases over longer periods of time therefore in the LETS it may be difficult for the respondent to recall how many times an event occurred but as respondents tend more towards general judgments over longer periods of time then this is more in keeping with using levels of agreement as the adjective (Schwarz and Oyserman 2001). In terms of the DOTS, although I felt that frequency may be easier to judge and that it was less personally critical of the trainer it did not make sense for all items to carry time associated adjectives such as "does a lot", "does a little". The other disadvantage in using adjectives involving frequency is that one of the categories therefore commonly chosen is 'often' and this can be difficult to interpret as it is largely based around how frequently a person normally experiences that event (Streiner and Norman 2008). I therefore opted to use adjectives based around agreement for both tools, although agreement could also be said to be subjective I felt that evaluation of training always has an element of subjectivity especially from the trainee's point of view.

5.1.2 Free-text comments

Alongside the items I also chose to include space for free-text comments. Although one could argue that a good tool should capture all components of teaching clearly it is difficult to capture idiosyncrasies that might be specific to an individual teacher or

events that may only occur rarely as this would not be useful for all trainers on every occasion that the tool is completed. Such idiosyncrasies can be captured by free text comments. Enabling respondents to make comment means that they are able to elaborate, as this can be powerful feedback in that it can either reinforce positive opinions or place greater emphasis on deficiencies. As well as expanding on their opinion respondents may also offer suggestions on how to change particularly in reference to deficiencies. Other research has also found that teachers commented that alongside ratings they found free text comments helpful (Stalmeijer, Dolmans et al. 2010) although this is not expanded on further. Free-text comments can also be used to give evidence for content validity by ensuring that no new themes arise from such comments(Cox and Swanson 2002). I therefore included instructions to complete a free text comment box to each tool that read,

> *Comments- please make these as specific as possible in order to inform your trainer about their teaching*

5.1.3 Instructions

The tool also contained some instructions for completion; these re-emphasised how the toolkit had been developed and why it needed trialling. It reminded trainees and peers to try and evaluate trainers as fairly as possible but also reassured them of their anonymity. The instructions suggested that the DOTS should be completed as soon as possible after the teaching episode to which it pertained. The LETS could be completed at any time but should be after a sustained period of training although I did not suggest a minimum for this.

The toolkit also collected demographic data about the person performing the evaluation in order to investigate whether there were any variables that made a difference. The decision about which demographic data to collect is discussed further in chapters 7 and 8.

5.1.4 Cognitive interviewing

In Chapter 3 I discussed the dilemma that arose in terms of when to perform cognitive interviewing. I opted to perform it prior to the Delphi process in order to try and ensure that all the attributes conveyed the same meaning to those that were involved in this consensus process. Performing the cognitive interviewing at this stage however meant that only issues with understanding were highlighted, whereas other potential

problems can be in retrieval and judgment (Conrad and Blair 1996) which could not be examined. It has also been shown that when students evaluate teachers some of the educational terminology was misunderstood but also that they often took into account factors other than those mentioned in a question; tended to use the high end of the scale and whilst they used the highest mark selectively they tended to use the next highest for varying reasons(Billings-Gagliardi, Barrett et al. 2004).

Due to the above I thought it was important to perform another round of cognitive interviewing. The items had also had their wording changed as a result of the Delphi process and so I felt it was important to check that the new wording was as clear as possible. As the items had already been refined by the first set of cognitive interviewing and the Delphi process I did not feel that I needed to repeat the same number of cognitive interviews therefore I only performed four interviews; one with a trainer and three with trainees. I selected trainees that were already training in endoscopy rather than those who were about to commence their endoscopy training (as in Chapter 3). This is because I wanted them to complete both evaluation tools about their trainer before being interviewed.

The trainees were asked to complete both aspects of the toolkit following an endoscopy training list with their current trainer. Similarly the trainer was asked to complete the tools with regard to his most recent training. All signed a consent form to agree to participate in the interview. Trainees were asked to leave blank the trainer's name in order to protect anonymity (as their trainer's consent had not been obtained). The three trainees and the trainer were then interviewed. Asking the trainee to complete the tool prior to the interview is a method to try and make sure their answers are as representative of how they actually felt and also it helps mirror how the tool will be used (Drennan 2003). In the interview they were initially asked to review the instructions and demographic data. They then went back through the evaluation toolkit using a 'think-aloud' process to justify their choices on the scale for each item. A selection of probes were pre-scripted and included

- Were you able to answer this question with ease?
- What does it mean?
- Could you suggest better wording?
- Was the scale appropriate?
- Any changes to the scale needed?

Each interview was conducted within the education centre of the trainee or trainer's place of work at a time of their convenience. The results of the interviews were then reviewed by myself and SC and changes made by consensus. Again if there was any doubt about the intended meaning we checked the comments made in the Delphi process and the original interview excerpts from Wells' interviews (as described in Chapter 3). The results of the interviews along with any changes that were made were then reviewed by JRB to try to ensure objectivity.

5.1.4.1 Results

The lapsed time between the interview and the list was anything from one to five days; each interview lasted 45 minutes to an hour. Each interviewee found the tool easy to use and there were no issues with the instructions. The trainer interviewed suggested that as well as asking about attendance at a 'Training the Trainer' course this question should also include other teaching qualifications or attendance at other teaching courses; this change was made. There were no other changes made to the wording of the instructions or demographic questions. The other main alteration suggested by the trainer was to change the wording at the beginning of the trainer self-assessment tool to read 'I as the trainer…' This is consistent with other self-evaluation tools that mirror student evaluation tools(Stalmeijer, Dolmans et al. 2010) and therefore was felt to be an appropriate change.

In terms of the items, for both the DOTS and LETS, trainees used a whole range of scores from strongly agree to strongly disagree. The cognitive interviews highlighted issues with six of the items in the DOTS and three of the items in the LETS.

One of the items that an issue was raised with was

'The trainer agreed objectives for the session'

One of the trainees disagreed with this attribute as the trainer did not do this at the beginning of the session; however did comment that their trainer had actually agreed objectives at the end of the previous session. In order to accommodate this, the following phrase was added in brackets 'either previously or at the beginning of the session'. Another item where further wording was added to make the item more explicit was in the item,

'The trainer checked that I understood questions and advice'

One of the trainees agreed with this item as they felt that the trainer had done this by observing; however the other two trainees did not consider observation as a method of checking; in order to highlight this option the phrase 'by observing or questioning' was added.

In the item that refers to the trainer ensuring the trainee knew the name and role of each member of the team, two of the trainees only concentrated on the concept of the name in the think aloud process whereas role is in fact the more important aspect (as that can change from list to list) therefore I opted to swap the order of the words 'name' and 'role' to try and place greater emphasis on the concept 'role'. This change was in keeping with primacy effect where people attend to the first item on the list more so than subsequent items (Duffy 2003). By changing the order of the items respondents are likely to place greater emphasis on the concept of the role of others in the team rather than just their name.

In Chapter 3 when I previously discussed the results of cognitive interviews I highlighted that examples can be both helpful and confusing. The potential to confuse was also highlighted in this round where in the item on specific skills teaching examples had been included. One trainee chose neutral for this item as he had not experienced the specific examples given but had received teaching on other areas; one of the other trainees discussed a similar situation but decided that this was still specific skills teaching and therefore agreed with the item. In order to highlight that these were only examples of skills teaching and not an exhaustive list the wording preceding the examples was changed to 'some examples of this might be'. In the LETS, an item that referred to the use of teaching aids had already had more examples added as a result of the first round of interviews. Interviewees however still only referred to the magnetic imager; although interviewees should consider this it was not the only example, I therefore moved the magnetic imager further down the list of examples to try and reduce its emphasis and to try and ensure that the toolkit was not too colonoscopy specific.

Other changes that were made to the LETS included adding 'during lists' to the item

'The trainer took advantage of opportune moments to teach'

This was added because two of the trainees when judging the trainer on this item referred to episodes that occurred during the list whereas the other trainee referred to

being called down to endoscopy at other times, for instance, because something interesting was happening. As this trainee's trainer had never done this the trainee disagreed with the item. As I was unsure whether this was included in the original intent of this item, I referred back to the original interviews performed by Wells (2010). In the excerpts that were contained under this node all the examples occurred during the list therefore this clarification was added to the item.

As mentioned in the previous chapter two items were unallocated at the end of the Delphi process. One of these was

'The trainer actively listened to me'

All three trainees and the trainer gave different interpretations to the meaning of this item; in the original interviews the meaning to this statement was also vague and included the use of non-verbal body-language and answering questions appropriately, both of which were mentioned by different trainees. Given the difficulty in ensuring this statement conveys the same meaning to all and the fact that there are already several items that examine dialogue between the trainer and trainee I decided to exclude this item from both tools. The other item that had not been allocated by the Delphi process was

'The trainer gave me opportunity to ask questions'

Interviewees felt that this could be included in either tool and therefore a decision was made on a pragmatic basis that as the LETS was to be completed less frequently than the DOTS then its length was less crucial therefore I opted to include this item on the LETS only.

5.1.5 Conclusion

Cognitive interviewing enabled me to pre-trial the tools within the workplace, both tools took less than five minutes to complete and no practical problems arose from trialling it. Trainees used different points on the Likert scales to complete the tool and felt able to complete every item.

The final tools can be found in Appendix 6, 7 and 8. Performing cognitive interviewing did lead to some further changes to the items in order to try and ensure that all items were interpreted in the same manner by all trainees. The majority of the items

remained unchanged but there were some exceptions as described above. In changing some of the items they became more prescriptive, such as limiting the item that referred to opportune moments to teach within an endoscopy list, whereas other items became more generalised, such as setting objectives either within this session or at the end of the last session. This item had resulted from a combination of two items in the Delphi process (Table 4-7) one of which referred to setting objectives at the beginning of a list and one at the end which were felt to be similar and amalgamated. On reflection these are actually slightly different concepts, which have been amalgamated to one item. The rewording of this item through the cognitive interview reflects this amalgamation better but one could argue that there has been a loss of information as setting objectives at the beginning and end of the list are slightly different concepts. In reducing the items with the Delphi process and trying to make the items interpretable to all I feel that there is some loss of detail but this is at the sacrifice of trying to ensure that the tools are usable.

The process of converting the items into a toolkit has contributed to the evidence for the Response Process particularly in relation to trialling the tool and performing further cognitive interviewing although this was limited by the number of interviews conducted. Over the next few chapters I shall discuss how evidence was gained for further sources of validity including the internal structure and response to other variables.

Chapter 6. Establishing Internal Structure and Reliability

So far the work done on the toolkit has aimed to contribute to its content validity and consider the response process but as mentioned in chapter 2 this is only two of several domains from which evidence for validity can be sought; other sources of evidence of validity are internal structure, relationship to other variables and consequences. In order to gain an assessment of these sources of validity the toolkit actually has to be used to evaluate trainers. To optimally investigate these sources of evidence it is important to consider how they could be assessed. In this chapter I wanted to expand particularly on the domain of internal structure, which includes the concept of reliability and consider what this entails in order to inform how I go about trialling the toolkit, which is discussed in the following two chapters. I discuss the concept of internal structure and the main methods by which this can be considered; I then look at the two main methods by which reliability can be examined, classical test theory and generalisability theory.

6.1 Internal structure

As previously mentioned internal structure is normally the most common source of evidence for validity when examining evaluation tools (Beckman, Cook et al. 2005). It refers to the statistical properties of the tool (Downing 2003) once it has actually been used or trialled. In order to consider internal structure further I have firstly considered evidence for the internal consistency of the tool, how the items relate and correlate with each, and then the reliability of the tool as a whole.

6.1.1 Internal consistency

Internal consistency is established using statistical tests that examine whether the items on a tool correlate with each other. Normally the aim of a scale is to attempt to measure a single construct or trait, in this case the ability of a trainer to teach a trainee. As all the items on the tool should be trying to measure this construct then every item on the tool should correlate with every other item on the tool (Beckman, Ghosh et al. 2004); this correlation is presumed to be the degree to which each item measures the construct under investigation.

One method by which internal consistency can be measured is by looking at item-total correlations (Streiner and Norman 2008). This is the correlation of an individual item with the total sum of the scale, having omitted the item under investigation from the total. The item itself needs to be removed as otherwise it would artificially elevate the degree of correlation. If all the items are meant to measure a single trait one would expect each of them to correlate with the total at a level of greater than 0.2.(Streiner and Norman 2008) or 0.3 (Field 2009). This process enables each item to be examined individually and assess to what extent it 'fits' with the rest of the tool.

If one wants to examine the internal consistency of the total scores then one method is to perform split-half reliability (Beckman, Ghosh et al. 2004). Split half reliability is performed by randomly splitting the items into two subscales. The total scores of the two sub-scales are then correlated. However this will not be the reliability of the whole tool, as it is only based on half the number of items and reliability always increases as item numbers increase(Streiner and Norman 2008). The reliability of the whole tool can then be calculated using the Spearman-Brown 'prophesy' formula(Streiner and Norman 2008). There are several issues with this as a method to calculate the internal consistency of the tool. One of these is that one needs to ensure that the items are randomly allocated to the two subscales but even then the reliability only represents the reliability given to that particular division of the scale and clearly there are many combinations by which the scale can be split which would result in slightly different reliabilities. Another disadvantage of this method is that it is not possible to examine which items are responsible for lowering the reliability (identifying those items that do not appear to be measuring the same construct).

An alternative to this is to calculate Cronbach's alpha, this approximates the average of all possible split-half reliabilities of a scale (Streiner and Norman 2008). One of the other advantages of Cronbach's alpha is that it can also be calculated with each item excluded. If Cronbach's alpha increases significantly once that item has been removed then it suggests that that item does not measure as much of the underlying trait as the other items and therefore can be considered for exclusion

There are however several points that must be taken into account when using Cronbach's alpha to show evidence of internal consistency. The number of items in the scale will affect the strength of the correlation; this is because the number of items forms part of the numerator for the calculation of Cronbach's alpha. This means that

as the number of items increases so will alpha, therefore it has been argued that it is an ambiguous marker of internal consistency as really it is a function of two parameters, the number of items and the average intercorrelation of the items(Clark and Watson 1995). Most researchers feel that Cronbach alpha is still an acceptable test but that note should be taken of the number of items and that those tools with more items require higher correlations as evidence of internal consistency (Streiner and Norman 2008).

Whilst above I have talked about the need for high correlations in order to show that a tool is measuring one underlying construct there is also a counter-intuitive argument that in fact too high inter-item correlations should also be avoided. The argument for this is that if two items correlate too highly and one is already included in the tool then the next item provides no extra information making the latter item redundant. Clark and Watson (1995) refer to this as the 'attenuation paradox' and argue that if items that correlate too highly are all included in a tool then this does not enhance construct validity and can in fact be damaging to the overall validity of the tool. They also argue that if the only driver behind the decision of which items to retain on a tool is item consistency then the construct being measured can become too narrow.

6.1.2 Factor analysis

Factor analysis is a method that examines whether subscales exist within a scale; it enables the researcher to know whether the scale is unidimensional or whether there are sets of items that go together or seem to stand apart (Rust and Golombok 2009). Factor analysis, rather than just looking at shared variance as a whole, tries to identify 'factors' (groupings of items) that are hypothetical constructs that can be used to explain the data (Rust and Golombok 2009).

Factor analysis can be either used in an exploratory or confirmatory way. Exploratory factor analysis refers to examining the data for underlying factors or domains when the investigator has not yet created a hypothesis of how a scale may be subdivided, whereas the converse is true in confirmatory factor analysis. As the toolkit does not yet consist of subscales and I have not created a hypothesis of what subscales may exist I will discuss just exploratory factor analysis. Exploratory factor analysis can be used for two main reasons; it can be used to look for underlying dimensions or domains of a measurement instrument(Floyd and Widaman 1995), or prove that no such domains exist and therefore demonstrate unidimensionality(Rust and Golombok 2009). It can

also be used to try and reduce the number of items by only including in a final tool those items that are maximally weighted (i.e. those items that displayed more of the shared variance in each of the domains) (Floyd and Widaman 1995).

Although exploratory factor analysis is very common in the development of a scale (Cortina 1993) it is a methodology that has been widely criticised(Floyd and Widaman 1995); this appears to be for several reasons. The data must meet several requirements which are not always adhered to; there are also many methods that can be used to perform factor analysis and the researcher has to make several choices along the way; the choices made can result in different results (Costello and Osbourne 2005). I will therefore examine some of these issues with factor analysis.

In order to perform factor analysis the data collected must satisfy certain requirements. One requirement is that there must be an adequate sample size(Field 2009). The size of this sample is related to the number of items; the larger the number of items being tested the larger the sample size required. There is no absolute number for sample size as data that loads very highly onto different factors requires a smaller sample (Costello and Osbourne 2005). Field (2009) states there must be a minimum of five to one cases to items; the Kaiser-Meyer-Olkin (KMO) and Bartlett's test of sphericity are also accepted ways of measuring the adequacy of the sample size(Field 2009). With smaller sample sizes the resulting factors must be considered with caution, as they may not be a true reflection of the data structure.

One of the decisions in factor analysis is to decide how many factors should be retained. Factor analysis will result in as many factors as there are variables (these variables in this case are items) however the amount of actual variance explained by some of these factors can be so small that it does not need to be accounted for, there therefore needs to be a method by which to decide how many factors to retain. The amount of shared variance within each factor is demonstrated by the size of a factor's eigenvalue; the larger the eigenvalue the greater the amount of variance that factor accounts for. One rule is to retain all factors that have an eigenvalue over one and this is termed Kaiser's criterion (Rust and Golombok 2009); this rule is derived from the fact that an eigenvalue of less than one is felt to be of little interest and likely due to error. Eigenvalues can be calculated by SPSS and the term itself relates to matrix mathematics (Field 2009). This is not always a foolproof rule; Rust and Golombok (2009) use the example that if factors with eigenvalues of 1.1 and 0.9 were found it would not make

sense to keep one and not the other as the difference in the amount of variance explained is very small; and studies have found that using Kaiser's criterion can both under and over estimate the number of factors retained(Floyd and Widaman 1995). An alternative method in determining how many factors to retain is to use a Cattell scree plot (Field 2009). In this method a scree plot is drawn with factors plotted against their eigenvalues; as the amount of variance explained by each factor successively decreases, the line seen on the graph is a downward slope; this line will at some point have a point of inflexion or an 'elbow'; the number of factors to the left of this point of inflexion is the number of factors to be retained(Field 2009). Again this method is not felt to be perfect and Costello and Osbourne (2005) suggest that if there is a cluster of factors around this point of inflexion then the factor extraction should be rerun for each of these numbers of factors and the investigator should then determine which statistically appears to have the best fit to the data. Regardless of how the number of factors is decided upon, Streiner (1994 cited in (Floyd and Widaman 1995) suggests that the factors retained should explain at least 50% of the variance, although Floyd and Widaman (1995) feel that factors should explain 80% of the variance.

In terms of the end result of factor analysis one wants to look at the factor loadings; these factor loadings are a gauge of the substantive importance of a given variable to a given factor (Field 2009) and the higher the factor loadings and the more variables contained within a factor, the more stable the factor design (Costello and Osbourne 2005)(by stability I mean that were it to be repeated with a different data set then the same results are likely to arise). Normally a loading of 0.3 is felt to be acceptable (Floyd and Widaman 1995) although this is also dependant on sample size and the smaller the sample size the greater the loading needed in order to be to be deemed acceptable (Field 2009). If items load onto more than one factor then this can be a sign of instability and that item should be reviewed. As well as the loading the size of the communalities should also be considered. The communality is the proportion of common variance seen within a variable (Field 2009) and one would expect variables within a factor to have communalities of at least 0.4(Costello and Osbourne 2005).

Litzelman et al (1993) used factor analysis to investigate the underlying structure of their tool, which is based on the Stanford Faculty Development Program framework. This is a framework developed for faculty development from which a 58-item toolkit was derived. The framework has seven categories and each category contained at least

seven items they also added an extra knowledge category which is not part of the SFDP but is included as a separate category in other tools. Using a large dataset the authors examined the internal structure of the tool using exploratory factor analysis. They split the dataset in half containing roughly 700 evaluations each and on the first half performed exploratory factor analysis using principal components analysis. In order to extract the factors they used three methods; firstly they extracted all factors with eigenvalues over one which led to a six-factor structure; they then preset the number of factors to be extracted at seven as the tool had been developed from a seven category model, and then eight factors to also include the knowledge category. Using these extractions they found within the six factor structure that two of the categories within their tool became spread over several factors and were not readily interpretable. Extracting seven and eight factors supported the hypothesis of this many factors in both cases although there were several factors that cross-loaded onto two factors and two items that were factorially ambiguous. In the seven-factor model the knowledge items collapsed into the 'promotes self-directed learning factor' and explained 77% of the variance. The second half of the dataset was then used to see if these factor structures were replicated using the same methodology. Only the seven-factor model retained consistent factors. They then also used this model to reduce the number of items to 25 from 58 original items by removing items that loaded poorly onto factors or loaded onto several factors or were ambiguous. This demonstrates how factor analysis can be useful to gain evidence for categories or domains within their scale and could be used to support an argument of how data should be used to feedback to teachers. It does however also demonstrate that for factor analysis to be effective large data samples are required.

6.2 Reliability

As discussed in chapter two reliability of the tool can be seen as a component of the internal structure. Reliability can be examined in many ways but is generally divided into classical test theory and generalisability; these two concepts will be discussed in turn

6.2.1 Classical test theory

In chapter two I introduced the concept that an observed score on a test is in fact composed of two components; a true score and an error associated with that

observation and it this concept of true score that forms the basis of classical test theory(Streiner and Norman 2008) as all formulas used to investigate data for reliability can be derived from this statement. This is because reliability attempts to represent how well the observed score estimates the true score and reflects the amount of error present in the resulting data from a test. Given this knowledge the overall equation for reliability is;

Equation 6.1. Equation for reliability

$$Reliability = \frac{true\ variance}{total\ variance}$$

$$Reliability = \frac{true\ variance}{true\ variance + error\ variance}$$

$$R = \frac{\sigma_S^2}{\sigma_S^2 + \sigma_e^2}$$

The first two lines of the equation were given in chapter two but it is expressed more mathematically in line three where σ_S^2 is the true variance of the subjects and σ_e^2 is the error variance. Obviously because the true score can never be known then neither can the true variance therefore there exists a variety of ways that this can be estimated.

When the subject of the test is being observed by another, such as in this case when the trainer is being observed by a peer or evaluated by a trainee, then the greatest risk to the reliability of the test is the inconsistency between observers or within observers themselves(Downing 2004). The amount of agreement between two raters could be calculated looking at the percentage agreement between them (Beckman, Ghosh et al. 2004) however this is an imperfect measure as some of this agreement will be due to chance agreement and therefore the apparent reliability of the test would be falsely elevated. There are some more sophisticated measures therefore which enable the comparison between two measures that account for chance within their calculations such as the kappa co-efficient(Cohen 1960) or the Pearson correlation (Streiner and Norman 2008).

One may not want to just look at the reliability between raters but also within raters themselves, for instance over time which can be performed by looking at the test-retest reliability, which can also be used to examine how the subject themselves may perform differently on different occasions.

The disadvantage of using classical test theory to calculate reliability is that it is only possible to look at one possible source of variance at a time and error is undifferentiated; for instance, if a reliability coefficient was calculated to account for variability due to different raters it would not quantify reliability over different cases or times in the same analysis. These other sources of variation are amalgamated into a single error term. This means that it can be difficult to tease out subtle causes for error; for instance raters tend to have different levels of stringency, commonly referred to as being either hawks or doves, but aside from this may also mark different cases differently, referred to as case specificity; it would not be possible using classical test theory to separate out these two possible sources of variance (Crossley, Davies et al. 2002). It is also not possible with classical test theory to compare the different contributions that sources of variance make to the overall test score(Crossley, Davies et al. 2002). Using Equation 6.1 classical test theory can only enable us to attribute error to one source of variation at any one time.

6.2.2 Generalisability theory

Due to these limitations in classical test theory alternative statistical models have been developed; one of which is generalisability theory. Streiner and Norman (2008) explain the fundamental difference between generalisability theory and classical test theory. In generalisability theory the true score does not exist but is referred to as the universe score. This is because the true score will always be affected by different test conditions, in generalisability theory the researcher must identify likely sources of error and the researcher must make a decision about which of these sources of potential error they want to generalise over; these sources of error are termed facets. Once they have decided this they then have a universe of possible test scores resulting from all the possible combinations of the test conditions; the score that would be found if all the scores were gained from all the possible combinations of the test scores is termed the universe score. This is subtlety different to the true score as it recognises that is still a condition of the test. Different sets of facets would result in a different universe score.

The main advantage of generalisability to the researcher is that once the different facets have been defined then it is possible to calculate the amount of variance due to these different facets simultaneously within the same statistical framework. It enables the researcher to perform a multivariate analysis compared to the bivariate analyses

performed in classical test theory. For instance this means that a variance component can be estimated for both rater stringency and case specificity within the same analysis as well as a variance component for the specific interaction between rater and case. Generalisability theory gives a further in-depth view to the error term and enables the researcher to quantify various factors that contribute to total variance. Variance components for different sources of variation can then be combined to create a G co-efficient; an estimate of reliability(Shavelson and Webb 1991).

In order to perform a generalisability analysis it is important to acknowledge certain properties that the facets possess. The facet that is the subject of interest, for instance the differing abilities of endoscopy trainers to teach, is termed the 'object of measurement'. All other facets are elements of measurement that alone or in combination could explain variance (Shavelson and Webb 1991). These facets can either be random or fixed. In order to decide whether a facet is fixed or random depends on the level to which the researcher wishes to generalise their results (referred to in generalisability theory as the universe of possible admissions). A random facet is one that is interchangeable and the sample used in the study is smaller than that of the universe of all possible samples of that facet, for instance trainees. A facet is said to be fixed if the conditions in which it is used in the test are the same as the conditions that the researcher wants to generalise to, for instance if the observers used in the study are the only observers that will ever be used then observers would be a fixed facet i.e. the sample used in the study represents the universe of all possible samples.

Facets are also described as crossed or nested. Crossed facets refer to those where every condition of each facet is repeated for every condition of another facet, for instance if every examiner examined every student on every case in a practical exam then all facets are crossed; whereas a nested facet is where certain conditions of a facet are only related to certain conditions of another facet, for instance if in the above example an examiner only examined on one case then examiners would be nested in case (Streiner and Norman 2008). Recognising the nature of a facet is important because it alters how the facet is inputted into an analysis and therefore can alter both the result and its interpretation. A fully crossed design is the most efficient variance component analysis and allows the variance contributed by each facet to be analysed fully but as Crossley (Crossley, Russell et al. 2007) points out that this is often not

practical and it is very difficult to set up a fully crossed test scenario in naturalistic settings therefore often study designs include nested facets. Each facet is notarised by the variance notation σ^2.

From a generalizability analysis it is then possible to look at the percentage that each variance component contributes to the total variance (Streiner and Norman 2008). Once the variance components have been calculated it is possible to calculate a G co-efficient, which reflects the reliability of the test under the conditions used.

The G co-efficient that results from a generalisabilty analysis is similar to a reliability co-efficient and represents the degree to which the results of the tool reflect all possible measures of the same construct (Crossley, Davies et al. 2002). The G co-efficient will always be lower than a reliability co-efficient as it takes into account all possible sources of variance at the same time(Crossley, Davies et al. 2002).

Once the sizes of different sources of variance are calculated in a generalisability study it is possible not only to calculate a G coefficient but also to use these variance components to make future decisions about what would reduce variance in future assessments or evaluations and therefore reduce potential error. One can then statistically hypothesise using the aid of a decision study how the test conditions could be altered to improve the reliability of the test results by mathematically modelling the G in different hypothetical test settings. As one would expect as the numbers of raters or cases that the test setting utilises increases the more likely that the results of the test are representative of all possible raters or cases. The size of the variance due to factors such as raters or cases is proportional to the size of this factor; for instance the more raters used the lower the variance due to raters.

6.2.3 Interpreting reliability

As mentioned previously both classical test theory and generalisability theory result in either a reliability or a generalisability coefficient which is given as a number between 0 and 1. Acceptable levels of reliability are discussed in chapter two; however as the coefficient refers to the reliability of the results of a test it reflects the ability of the test to accurately differentiate between individuals. It is difficult to know what this means for an individual's score and how accurate their score is. One way of using the reliability coefficient to interpret the accuracy of one individual's score is to calculate the standard error of measurement (SEM)(Streiner and Norman 2008). The SEM is an

absolute measure and quantifies the precision of an individual's score (Weir 2005 cited in(Streiner and Norman 2008)). It can be calculated using the equation below where σ_x is the standard deviation of the observed score and *1 – R* is one minus the reliability coefficient (Equation 6.2)

Equation 6.2. Equation used to calculate the standard error of measurement (Streiner and Norman 2008)

$$SEM = \sigma_X\sqrt{1 - R}$$

Using the SEM a confidence interval can be calculated using the following equation

Equation 6.3. Equation used to calculate confidence interval (Streiner and Norman 2008)

$$X_O \pm Z(SEM)$$

Where X_O is the observed score and Z is the value from the normal curve associated with the desired confidence interval; for a 95% confidence interval the value of Z is 1.96. A 95% confidence interval indicates that the true or universe score for that individual lies between the upper and lower values(Field 2009).

6.3 Summary

In this chapter I have discussed the various ways that both the internal structure and the reliability of a tool can be evaluated. Although generalisability has significant advantages in terms of being able to directly compare sources of error; it is not always possible in terms of data collection. For instance Crossley et al (2002) discussed that although a fully crossed generalisability design is ideal, particularly when testing a tool in the real world this is not always possible and the study starts to lack ecological validity (Cohen, Manion et al. 2007) as the study design moves away from the setting in which the tool will actually be used. If a naturalistic design is fully nested it could be argued that there is not advantage of generalisability over classical test theory as it is not possible to separate out the sources of variance and therefore only one source can be examined. This information can then be used when considering how best to trial the toolkits; this is discussed in the next two chapters.

Chapter 7. Peer Evaluations

The DOTS was designed so that it could be used by peers, trainees and trainers as a self-assessment exercise in order to give feedback to trainers regarding their teaching performance. In this section I will discuss the results of validating the DOTS with peers with a particular focus on assessing the internal structure and the reliability using some of the methods discussed in the preceding chapter.

7.1 Peer Evaluations

The toolkit has been developed in order to give feedback to local trainers working within their local units therefore ideally in order to gain an assessment of the internal structure and the reliability of the DOTS it should be trialled within these units. To assess between (interrater) and within (intrarater) rater reliability I would need two ratings from each rater on two occasions. The logistics of this would prove difficult; a peer review would require finding a further endoscopist to be present at training lists in order to complete the tool. Furthermore this would result in only one peer evaluation and decisions would need to be extrapolated from a single observer regarding the number of peers needed for optimum reliability. In already busy endoscopy units and consultants' timetables, having one endoscopy peer let alone more would have required massive resources and planning which were beyond the scope of this project. I therefore opted to use the JAG 'Training the trainer' courses in order to gain multiple peer evaluations. Using the training courses would allow me to gain evidence of its reliability without such expense and could be then be used as an argument for trialling the DOTS with peers in local units at a later date

The selected 'Training the trainer' courses were run at regional training centres throughout the UK. The courses are aimed at 'experienced endoscopists who are involved in teaching and training within the endoscopy service'(JAG 2012) and course participants tend to be those that are already involved in training within their own units and are looking to improve their training skills. The courses are not compulsory but are recommended for those involved in training, and training leads within units are expected to encourage their trainers to attend.

The courses were two days in length and there were six participants per course with two to three faculty members who were experienced trainers. Day one of the course

involved group discussion and practice on mannequins and the content focused on adult learning theory, skills teaching techniques and objective setting (JAG 2012). The content of this day may have varied slightly from centre to centre but was largely based around a curriculum set by JAG. On day two of the course each course participant had the opportunity to teach on a single case using a real patient. This was observed by a member of faculty within the endoscopy room and then the rest of the course attendants and faculty observed the case via video link to a seminar room, therefore this meant that several people observed each case in its entirety. I therefore wished to use this opportunity to gain multiple peer evaluations for a teaching case which it would be difficult to achieve within the endoscopy unit. I believe that the course attendants were valid peers as they had opted to attend the course, were likely to already be teaching within their own unit and have all attended day one of the course.

7.2 Study design

In order to analyse the reliability of using peer evaluations for the DOTS I decided to use generalisability theory. The reason for using generalisability in this setting was because it enabled the examination of several different possible sources of variance within the one analysis. A detailed description of how I performed the analysis is discussed within the methodology section; however I felt that it was important to consider how the study would be designed and the analysis that would be performed prior to collecting the data

The purpose of a generalisability study is to try and provide as much information about the sources of variation in the measurement as possible in a single multivariate analysis (Shavelson and Webb 1991). In order to do this I needed to identify the 'object of measurement' which in this case was the trainer teaching on each case as it is the difference in results due to a trainer's ability to teach that I wanted the DOTS to measure. The other facets (elements of measurement that alone or in combination can explain some of the variation or could lead to error) I initially identified are shown in Table 7-1.

Table 7-1. Table of types of variation when trialling the DOTS with peers

Source of variability	Type of variation	Variance notation
Trainers (t)	Universe score variance (object of measurement)	σ^2_p
Peer reviewers (r)	Potential source of error related to whether reviewers are natural hawks or doves	σ^2_r
Cases (c)	Potential source of error related to the nature of the endoscopy case	σ^2_c
Items (i)	Potential source of error related to the items on the tool	σ^2_i
Trainees (s)	Potential source of error related to differences in training different trainees	σ^2_s
Trainer: peer interaction (t:p)	Potential source of error related to inconsistencies between a reviewer and a particular trainer's behaviour	σ^2_{tp}

All the facets, except items, were random as, in the future, I wanted the results to apply to any possible case or peer. Cases were nested within trainer as each trainer taught on a unique case. As the procedure is uncomfortable, invasive and carries some risk it would not be possible to ask a patient to undergo two endoscopies purely for the purpose of teaching therefore there was no crossover of cases between trainers; wherever the toolkit was trialled or will be used in the future this will always be the case. Additionally on the TTT courses each trainer only teaches on one case; therefore it was not possible to compare scores by peers for different cases and investigate how much variance in scores is caused by case specificity. Different cases are likely to provide different teaching opportunities and are therefore likely to be a source of variation however it is not possible to quantify it as a separate source of variation within this study. The other disadvantage of only teaching on one case is that this means that the trainer only trains one trainee. This means it is also not possible to look at variance due to differences between training different trainees. Clearly the personality and skill of a trainee will alter the dynamic of the training and different trainers are likely to interact with different trainees different therefore there would also a trainer: trainee interaction. These are disadvantages of this study design and using the TTT courses to trial the tool; however it does allow us to examine the reliability of the tool between different peers, which as discussed above would have

been difficult within a naturalistic setting. Items are included as a facet in the above design however because I have chosen specific items they are fixed as in the future I would only wish to generalise over the same items.

Using these facets it was possible to estimate several variance components; note in the table there is a facet for trainer/peer interaction as well as separate facets for trainer and peer. This is because not only will reviewers have a natural tendency to either be stringent or lenient in their ratings, there will also be variance due to the way they react to different trainers; this can be referred to as trainer specificity and reflects the fact that a reviewer will naturally respond to different trainers teaching behaviours in different ways. I have not included in the table a trainer/case interaction or reviewer/case interaction, undoubtedly these do exist but as mentioned above it is not possible to measure variance due to case in this design and therefore it is not possible to separate out variance for these interactions either.

7.3 Methodology

7.3.1 Data collection

All the courses running from November 2011 to March 2012 were identified through the JETS website (JAG 2012). The lead faculty member for each course was emailed to ask if they would be willing for the course on which they were teaching to be used to trial the tool. The email detailed the study and a copy of the participant information letter, consent form and the DOTS tool were attached. Once the lead faculty had replied to indicate their agreement I liaised with the course administrator who sent an information letter and the consent form to all those attending the course. Course attendants and faculty were asked to complete a consent form if they were willing to participate.

On day two of the course, participants were asked to complete a copy of the DOTS (Appendix 6) for each of the training episodes they observed. Normally verbal feedback is given to the trainer by the member of faculty who was in the endoscopy room with them, I asked for the tools to be completed prior to this feedback so that it did not alter others' opinions. This process was administered by me for one of the courses, and the course administrator for the remainder. Trainers did not receive the results of the completed tools.

7.3.2 Data analysis

Each course and trainer (including those trainers who were attending the course and those acting as faculty) were given a unique identifier in order to anonymise the data. The data was then entered into a database on SPSS version 14 (SPSS 2005). On several of the courses participants were explicitly told not to handle the scope themselves therefore it was not possible to evaluate items 12 (*The trainer demonstrated a procedure where necessary*) and 13 (*The trainer intervened in a timely fashion (either at a predefined time or if I am struggling*)) therefore these were excluded in the analysis.

The analysis of the data occurred in two stages, stage one involved exploring the data to examine variable frequencies, item distributions and correlations. These were then used to inform decisions as to how the generalisability analysis should be performed in the second stage.

7.3.2.1 Exploratory data analysis

Number of courses, trainers and evaluations performed were noted. Demographics of respondents were recorded and response frequencies were analysed. The data was explored to look at the general spread of data and review the way in which the tool had been used; a Q-Q plot was drawn and z scores calculated (Field 2009 pg 139) to explore for normality. An overall mean and standard deviation for the all evaluation scores was calculated as were individual means and standard deviations for total scores for each trainer. To explore potential differences between courses mean, standard deviation, and 95% confidence intervals were calculated. A one-way ANOVA was also performed to see if there was a significant difference between scores on different courses.

Item analysis was performed to ensure that a good spread of responses on the Likert scale had been utilised for each item. If any items had been scored the same by all participants it would be considered for removal as it would not be helpful in differentiating between trainer ability (Rust and Golombok 2009). This is also true of reviewers, if they scored every trainer the same this may be because they felt the trainer's abilities were the same but if this consistently occurred then one could argue either that that the reviewer does not help differentiate or that the toolkit does not, in its current format, enable the reviewer to distinguish between trainers. The data was also reviewed to ensure that no reviewer had given a trainer the same score for every item.

7.3.2.1.1 Missing data

I also reviewed the data to quantify the amount of missing data. The amount of missing data by item was examined as was the amount of missing data overall. Less than 15% missing data per item was acceptable for that item to still be included in the final analysis (Fox-Wasylyshyn and El-Masri 2005). Scores that included missing data were initially reviewed with missing data excluded but in order to explore whether this made a difference case mean substitutions were inputted as this is acceptable at the item level when less than 30% of the data is missing (Fox-Wasylyshyn and El-Masri 2005).

7.3.2.1.2 Internal structure

In order to examine the internal structure of the tool, item-total correlations were calculated with the item in question deleted (item–corrected total correlation). Cronbach's alpha was calculated for the tool overall and the alpha if item deleted for each item was also calculated.

Factor analysis was performed to examine the underlying structure of the data to see if there was a single domain or several different domains. If clear domains were identified then it may later guide how results of the tool are fed back to trainers. The data was examined to check it met the criteria required to perform factor analysis (discussed in Chapter 6), this included the calculation of the item to case ratio, the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy and Bartlett's test of sphericity.

Factor analysis was performed using principal axis factoring as this separates unique variance in an item from shared variance between items and only analyses the shared variance(Costello and Osbourne 2005). It was performed using an oblique rotation using direct oblimin. This was chosen as it is the preferred method if a degree of correlation between items is expected(Field 2009). Given that I expected the tool to measure training effectiveness if different factors existed I would expect them to correlate to some degree as trainers who are good trainers overall would likely score 'better' in more than one domain.

Both the Cattell 'scree' plot (Rust and Golombok 2009) and Kaiser's criterion of retaining factors with eigenvalues over one (Field 2009) were used to try and determine the number of factors within the data and the number of factors were extracted accordingly. The strength of factors was reviewed using the guidance

discussed in chapter 6 (e.g. size of loading and cross loading onto different factors). Cronbach's alpha was also calculated to determine the reliability of items within each of the domains identified using factor analysis in order to consider the internal consistency of these domains.

7.3.2.2 Generalisability analysis

Following the above initial exploratory analysis it was then possible to perform generalisabilty analysis. One of the statistical assumptions made when performing a reliability analysis is that items must be locally independent of each other (Downing 2004) i.e. the units of measurement used should not correlate too closely. In the DOTS the items are nested within the tool and they are likely to correlate with each other (to be checked using Cronbach's alpha) therefore this assumption of local independence may be violated. In this event analysis must occur at the level of the case, items was therefore excluded as a facet as the individual items (or domains) are unlikely to be locally independent of each other.

In order to run the generalisability analysis I needed to decide whether to use the total score or the mean total as the dependent factor. Using the total score would mean that only complete evaluations could have been used, therefore any evaluation that contained any missing data would have been excluded, thereby reducing the data set. Using the mean of available scores for each evaluation allowed the whole data set to be used. I opted to use the mean scores for each completed DOTS as the dependent variable. This means that when the standard error of measurement is calculated this will also be based on the mean of the total rather than the total score; in order to accommodate this the SEM was multiplied by the number of items.

Generalisability was performed using the General Linear Model with the MINQUE model selected as this gives the best estimates when dealing with an unbalanced design (Crossley, Russell et al. 2007). The data was unbalanced as not all trainers received the same number of evaluations.

Initially I hypothesised that the main causes of variance would be those shown in Table 7-1 and include differences in the trainer's ability to teach i.e. the true variance and the peer reviewer's variability in the scores they gave, whether as a reviewer they had a natural tendency to be a hawk or a dove. I therefore ran a generalisability analysis with both trainer and peer reviewer as random factors using a full factorial design meaning

that the variance due to the interaction between individual trainers and individual reviewers would also be accounted for. Variance components were analysed using the General Linear Model function in SPSS as described by Crossley, Russell et al (2007). The output from SPSS reported variance components for each facet. These were converted into percentage of total variance, as this made it easier to compare the amount of variance accounted for by each facet. As using MINQUE does not report degrees of freedom the analysis was re-run using ANOVA sum of squares type III as suggested by Crossley, Russell et al (2007).

Using the variance components G co-efficients were calculated for varying numbers of reviewers using Excel (Microsoft 2007). The equation used to calculate the G co-efficient (Equation 7.1) mirrors the equation for reliability discussed in chapter two. The denominator is the sum of all possible causes for error variance; in line two of the equation this has been broken down into its component parts.

Equation 7.1. Equation for generalizability coefficient

$$Generalisability\ co-efficient = \frac{Variance\ component\ for\ trainers}{sum\ of\ all\ variance\ components}$$

$$Generalisability\ co-efficient = \frac{Variance\ component\ for\ trainers}{(Var\ comp\ trainers + var\ comp\ reviewers + var\ comp\ reviewer{:}\ trainer\ interaction)}$$

One advantage of generalisability, as previously discussed is that it enables the prediction of G co-efficients for varying evaluation circumstances, in this case differing number of reviewers. These G co-efficients are calculated in the same way as above but the differing numbers of reviewers need to be accounted for. This is done by considering which variance components will be affected by changing the number of reviewers. These variance components are divided by the number of reviewers under investigation; as the amount of variance attributable to these facets will be reduced if the numbers are increased. For instance to calculate the G co-efficient for two reviewers the above equation would be utilised but the variance component for reviewers and the variance component for the reviewer: trainer interaction would be divided by two (Equation 7.2)

Equation 7.2. Generalisability co-efficient when want to calculate for effect of two reviewers

$$Generalisability\ co-efficient = \frac{Variance\ component\ for\ trainers}{\left(Var\ comp\ trainers + \left(\frac{var\ comp\ reviewers}{2}\right) + \left(\frac{var\ comp\ reviewer:trainer\ interaction}{2}\right)\right)}$$

The standard error of measurement (SEM) was also calculated for each G co-efficient and the 95% confidence intervals using the equations in chapter 6 (Equation 6.2 and Equation 6.3)

In the next stage I performed a second generalisability analysis. This was performed using the same principles and methods as above but rather than selecting the sources of variance, by using what I hypothesised to be the main sources of variance, all possible sources of variance were inputted; this therefore included, alongside trainer and reviewer, the course, specialty of trainer, speciality of reviewer, and role on course i.e. whether course attendant or faculty. It was not possible to use a full factorial design as some of the interactions were nonsensical therefore all sensible interactions which gave positive variance components were kept. When the analysis was run if any of the variance components were negative indicating a poor fit of the model (Shavelson and Webb 1991) the most negative factor or interaction was removed and the analysis re-run; this occurred until all variance components were positive. A type III ANOVA was used to estimate degrees of freedom and to check the fit of the data to the model. In order to ensure that the analysis was not affected by using the mean score the analysis was re-run using total scores. The analysis was also performed excluding the course I had attended (Course 100) to explore for observer effect.

7.4 Results

7.4.1 Demographics

I contacted ten courses; eight of these agreed to participate in the study. These courses were held at six different training centres. 189 evaluations were collected in total; these were completed by 58 different peers with each peer completing one to five evaluations each. Forty-five trainers were evaluated and these received from one to ten evaluations each.

Overall there were seven surgical course attendants, one nurse practitioner and one general practitioner; the remaining were all gastroenterologists. All faculty were gastroenterologists. All procedures performed were colonoscopies.

7.4.2 Exploratory data analysis

7.4.2.1 Exploratory data analysis by trainer

Initial descriptors of the data are shown in Table 7-2. The mean total score for an evaluation was 63.3 (out of a possible 85) with a standard deviation of 8.6. There was a good spread of scores with a range from 31 to 85.

Table 7-2. Descriptive statistics of evaluation scores for DOTS peers

Descriptor	Statistic
Mean (standard error)	63.32 (.670)
Median	64
Variance	74.62
Std. Deviation	8.64
Minimum	31
Maximum	85
Range	54
Interquartile Range	9.
Skewness	-.405
Kurtosis	1.77

A histogram of the scores is shown in

Figure 7-1; this again shows that there was a good spread of scores. The histogram shows that the data was close to a normal distribution although there is a slight shift to the right of the scale (also shown by the negative skewness in Table 7-2). To more formally assess for normality a Q-Q plot was drawn (Figure 7-2); this plots the cumulative probability of a value against the expected cumulative probability (signified by the straight line). It is therefore possible to see that although there are a couple of outliers the data appears to map fairly closely to the expected value line suggesting normality. A z-score for skewness and kurtosis were calculated (by dividing each by its standard error (Field 2009)). The z score for skewness was 0.215 which was not

significant at the $p<0.01$ level i.e. the skew of the data was not significantly different from a normal distribution. The z score for kurtosis was 4.72 which was significant even at the $p = 0.001$ level (Field 2009)pg 139; suggesting that the degree of kurtosis (amount that the data clusters in the tails of a frequency distribution) was significant, however Field (2009) states that these values can be significant with very small deviations from the norm. Given that the z score for skewness was not significant and the Q-Q plot looks relatively normal I decided that the data could be treated as having a normal distribution for subsequent analysis. Normal distribution is important because if all the results are clustered at one end then statistically the tool will have better reliability but in reality will not be very good at actually distinguishing between levels of training proficiency. Although all the data is towards the top end of the scale it still has a normal distribution curve.

Figure 7-1. Histogram showing spread of evaluation scores

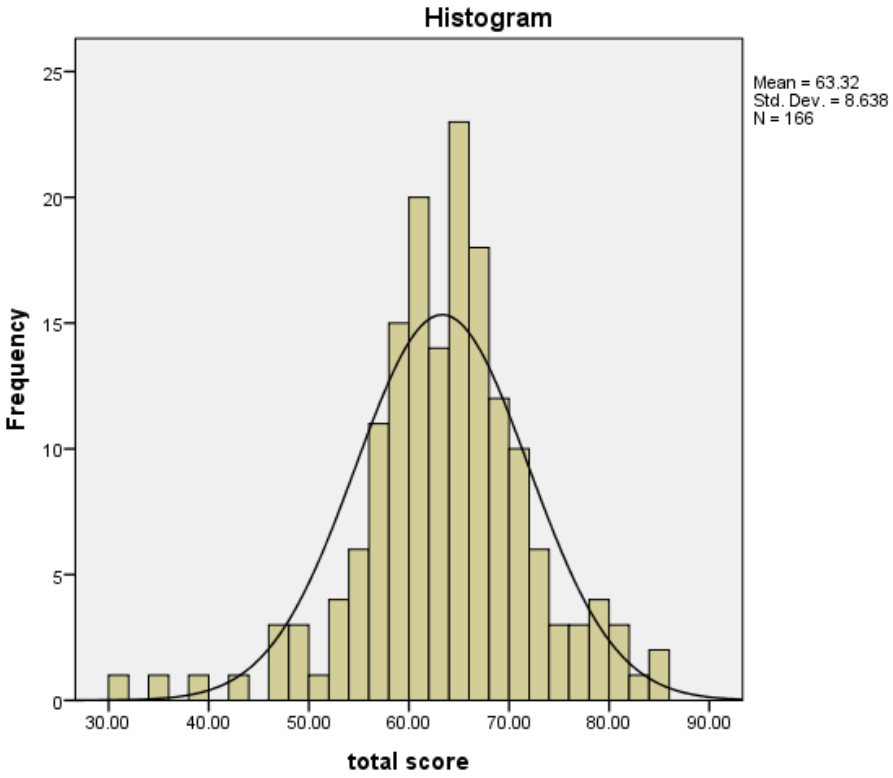
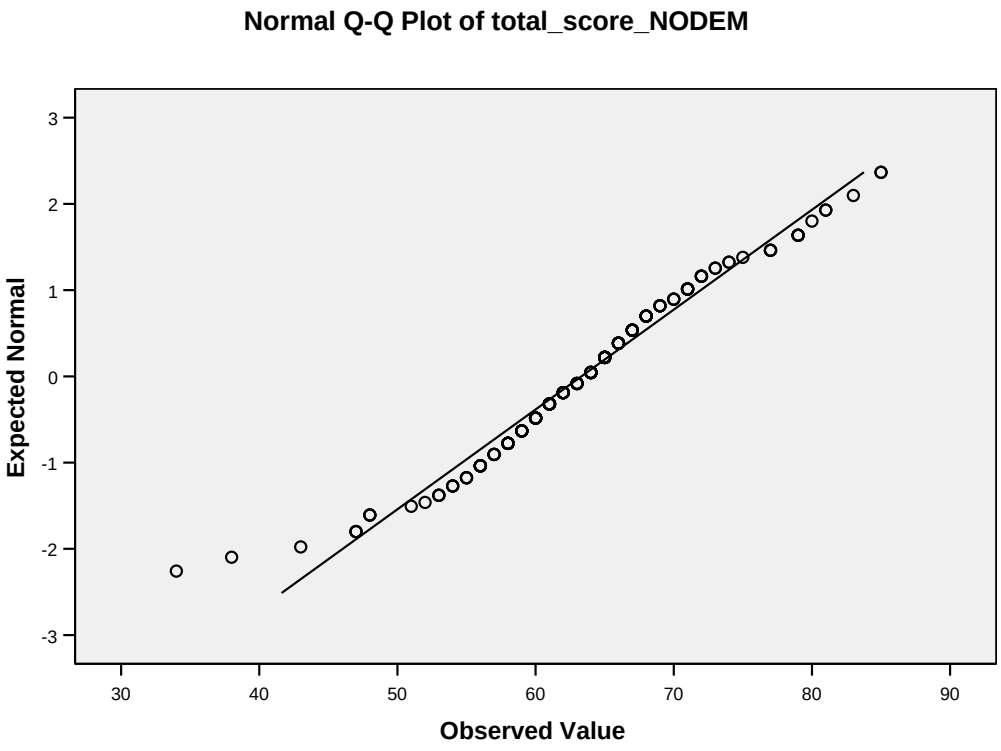
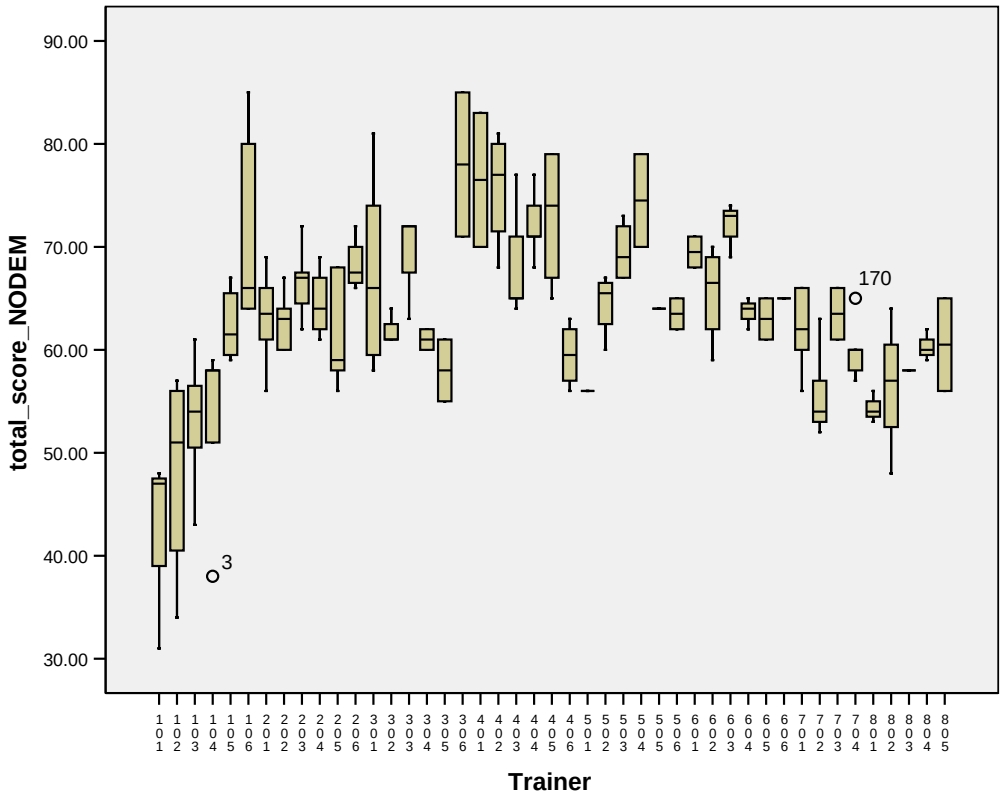


Figure 7-2. Q-Q plot to test for normality of peer evaluations on the DOTS



A box plot is shown in Figure 7-3 to demonstrate the spread of scores that each trainer received. The line through each box indicates the median score and the box itself represents the interquartile range. The 'whiskers' (the lines extending from each box) represent the highest and lowest scores, excluding those that are outliers (more than two interquartile ranges from the mean). As can be seen from this boxplot there was a good spread of scores between trainers, however, there was also a large spread of scores for some individuals; for instance trainer 106 received scores which ranged from 64 to 85. This spread of scores indicates that different reviewers had different opinions about a trainer's skill. Also there appeared to be differences in the ranges of scores given dependent on the course the trainer attended (course is denoted by the first number in each unique identifier for trainer), for instance in course one there was a large spread of scores between and within trainers whereas on course two the range was much smaller.

Figure 7-3. Boxplot to show spread of scores by trainer when evaluated by peers



Differences in courses was investigated further by calculating the mean, standard deviation and 95% confidence interval for each course (Table 7-3); as some of the confidence intervals for courses do not overlap this suggested there were differences

between courses. A one-way ANOVA was performed and demonstrated a statistically significant difference between course scores (Table 7-4).

Table 7-3. Descriptive statistics for peer data by course

Course	Mean	Standard deviation	95% Confidence Interval	
			Lower	Upper
100	56.6	11.3	52.7	60.6
200	64.4	4.2	62.9	66.0
300	66.3	8.3	62.3	70.3
400	70	7.4	62.2	73.5
500	66.6	5.6	63.3	69.8
600	66.6	4.3	64.2	69.0
700	59.8	4.6	57.4	62.1
800	57.7	4.8	54.6	60.7

Table 7-4. One way ANOVA comparing effect of course

	Sum of squares	Degrees of freedom	Mean square	F	Significance
Between groups	3820.6	7	545.8	10	.000
Within groups	8491.5	158	53.7		

7.4.2.2 Exploratory data analysis by item

Examining the items showed that a range of scores had been used for all items as can be seen in Table 7-5 and therefore each item was helpful in discriminating between trainers. If a range of scores had not been used for any one item then that item would have been considered for removal.

Table 7-5. Item analysis table of DOTS peer data

Data groupings	Strongly disagree	Disagree	Neutral	Agree	Strongly agree	Missing	Skewness
Q1	2	3	15	112	50	7	-1.205
Q2	1	11	64	81	18	14	-0.245
Q3	1	7	35	109	31	6	-0.704
Q4	0	23	36	103	27	0	-0.567
Q5	1	31	46	84	26	1	-0.322
Q6	2	17	31	104	34	1	-0.808
Q7	1	39	57	70	21	1	-0.084
Q8	3	33	32	94	26	1	-0.566
Q9	1	31	54	89	14	0	-0.364
Q10	1	24	41	99	21	3	-0.569
Q11	3	19	27	102	38	0	-0.871
Q14	1	5	12	123	45	3	-1.097
Q15	1	13	27	104	44	0	-0.820
Q16	4	31	51	79	23	1	-0.366
Q17	0	18	34	103	30	4	-0.606
Q18	0	11	14	114	46	4	-0.949
Q19	1	16	46	87	35	4	-0.458

Although a range of scores was used for each item it is possible to see that this was clustered to the right of the scale. This suggested that at an individual item level the distribution of ratings did not follow a normal distribution. The skewness of the data was therefore calculated for each item (and is shown in the right hand column in Table 7-5). All items were negatively skewed; demonstrating that the data is clustered to the right of the scale for all items.

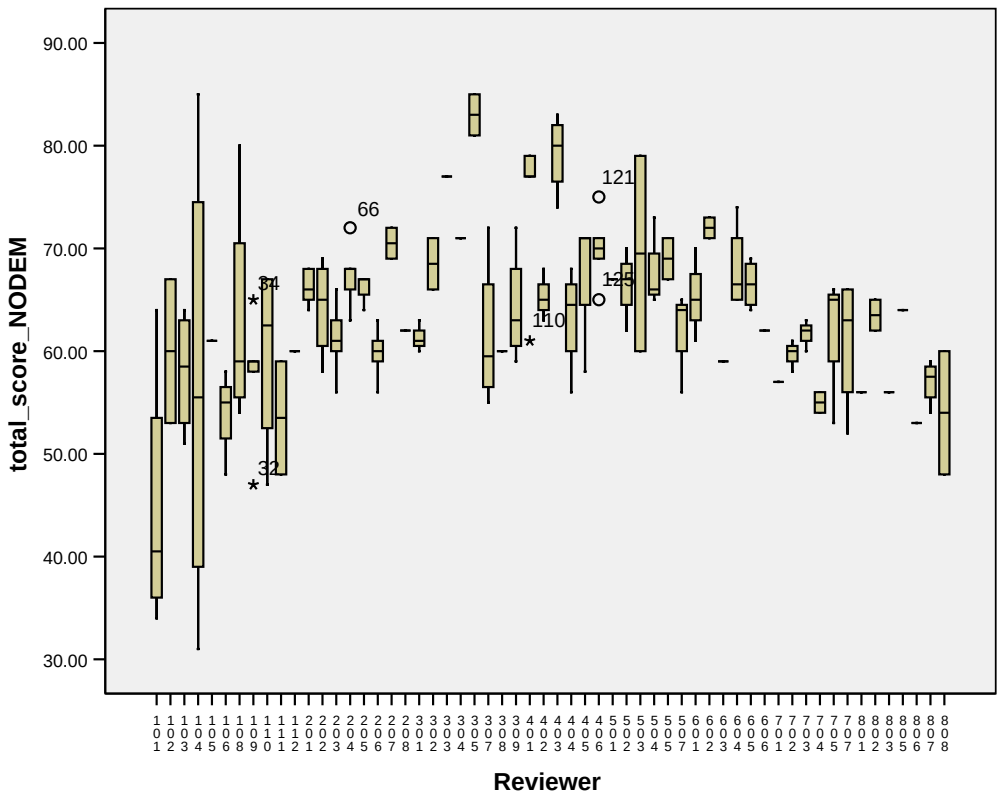
7.4.2.3 Exploratory data analysis by reviewer

A boxplot of the total scores given by reviewers is shown in Figure 7-4. From this it is possible to see that most reviewers awarded different total scores to different trainers for instance reviewer 104 used the entire range of scores giving both the highest and lowest score. However some trainers appear to have always given the same score (indicated by a single line on the boxplot). The data for these reviewers was reviewed

further to investigate why they had appeared to give the same total score. One reason is that some reviewers had only completed one evaluation. Another reason is that total score was calculated on SPSS by adding up the individual scores; if on SPSS there is any missing data then it is not able to calculate the total score using this function therefore many of the reviewers appeared to have only one total score as they had not scored all items. Only one trainer (501) had given the same total score to two different trainers; however on review of the score breakdown the reviewer had scored the two trainers differently on different items.

Two reviewers had given a trainer the same mark for every item. Reviewer 104 awarded trainer 106 'strongly agree' for every item; however this may be because this trainer had performed very well as the total scores awarded by this reviewer to other trainers ranged from 31 to 64. Trainer 305 also awarded 'strongly agree' for every item for more than one trainer (trainers 302, 303, 304 and 306) which may suggest that this reviewer was not using the tool to discriminate between trainers, however this reviewer did use varying scores for trainer 301.

Figure 7-4. Box plot to show range of scores given by each reviewer



7.4.2.4 Missing data

1.5% of the data was missing overall. The amount of missing data per item is shown in Table 7-5. For some items there was no missing data but for Q2 (*The trainer ensured the trainee knew the name and role of each member of the endoscopy team before a training encounter so that the trainee was supported*) 7.4% of the data was missing. For the rest of the items no more than 3.7% of the data was missing and for half the items less than 1% of the data was missing. This missing data will have affected the range of scores given by different reviewers to different trainers as some of the differences between scores will be due to missing data. Mean scores were then generated for each missing data point using the missing data function on SPSS, totals were then recalculated and the above analysis repeated (mean 63.6, standard deviation 8.98 range 31 to 85). The two sets of data were comparable suggesting missing data had not adversely influenced the spread of scores.

7.4.3 Internal Structure

Cronbach's alpha for the whole tool was 0.895 demonstrating that the tool showed high internal consistency. Table 7-6 shows the corrected item-total correlation; this demonstrates how well one item's score was internally consistent with the composite score from all other items that remain. Corrected item–total correlations varied from .416 to .652 which is above the cut-off for concern of 0.3 (Field 2009). Although Cronbach's alpha was high this was likely in part due to the fact that there were a large number of items, therefore to ensure each item does contribute to the apparently high internal consistency 'Cronbach's alpha if item deleted' was also calculated for each item; this is shown in the second column of Table 7-6. The removal of any item would decrease the overall alpha meaning that all the items contribute to the overall high correlation within the tool (rather than this just being due to the number of items) and all item- total correlations were strong.

Table 7-6. Item-corrected total correlations for the DOTS when completed by peers and the alpha for the whole tool if that item were deleted.

Item	Corrected Item-Total Correlation	Cronbach's alpha if Item Deleted
Q1_agreed objectives for the session	.524	.890

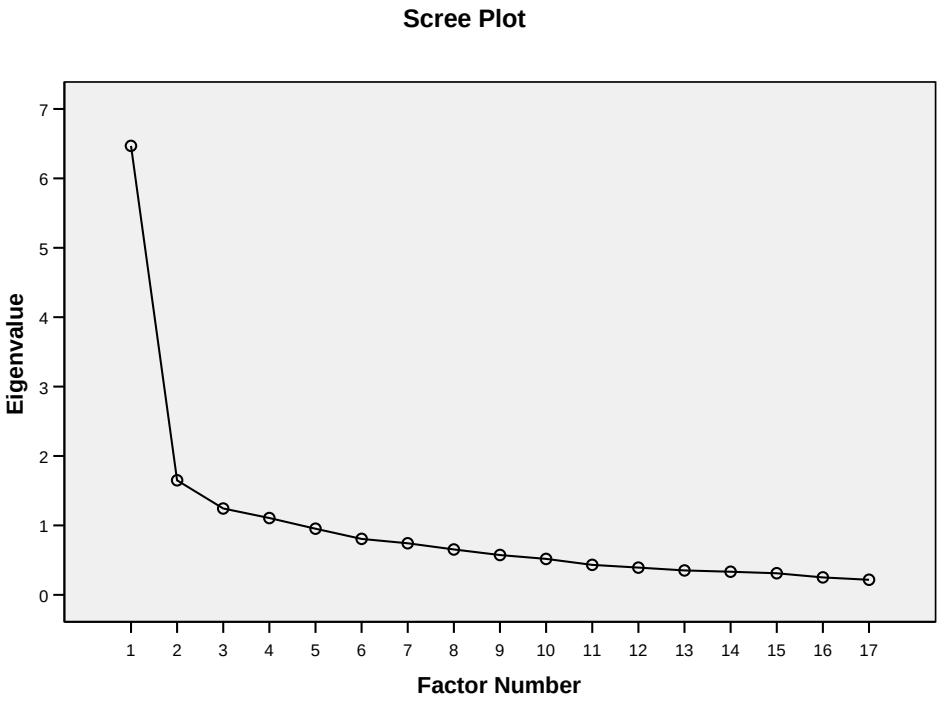
Q2_ensured the trainee knew the role and name of each member of the endoscopy team	.416	.893
Q3_agreed and applied the ground rules	.594	.888
Q4_questioned the trainee	.638	.886
Q5_provided explanations and descriptions	.560	.888
Q6_used a mixture suggestions, prompts, solutions and instructions	.589	.887
Q7_checked the trainee had understood instructions and advice	.582	.887
Q8_used an appropriate quantity of dialogue	.585	.887
Q9_asked the trainee to show where he or she was struggling	.479	.891
Q10_gave specific skills teaching	.590	.887
Q11_ did not overburden the trainee with too many tasks	.447	.892
Q14_allowed the trainee reasonable time to carry out a procedure	.571	.888
Q15_ always ensured the patient was comfortable and safe	.463	.892
Q16_encouraged the trainee to communicate appropriately with the patient	.465	.892
Q17_helped the trainee to assess if objectives for the session had been achieved	.592	.887
Q18_reinforced positive aspects of the trainee's performance	.652	.886
Q19_identified aspects for the trainee to develop and improve	.571	.888

7.4.3.1 Factor analysis

Principal axis factor analysis was performed. One hundred and sixty-six complete evaluations were collected giving a case to item ratio of 9:1. The KMO was .862 and Bartlett's test of sphericity was less than p .05 level of significance; suggesting the data

was suitable for factor analysis. A scree plot for the data is shown in Figure 7-5; the scree plot plots eigenvalues on the y-axis against factors on the x-axis. The number of factors to be retained is determined by the number of factors (i.e. the x-axis) to the left of the inflection point. From this it is possible to see that the scree plot suggested that most of the variance was explained by a single factor; this one factor would explain 38.1% of the variance.

Figure 7-5. A scree plot of peer data to determine the number of factors to be extracted



When extracting all factors with eigenvalues over one (Kaiser's rule), four factors were extracted. These four factors explained 61.5% of the variance. The factor matrix and pattern matrix are shown in Table 7-7 and Table 7-8; for ease of interpretation all items with loading of less than 0.3 were suppressed. The factor matrix (Table 7-7) demonstrated that although there were four factors with eigenvalues greater than one all but two items loaded highest onto factor one, but nine loaded onto more than one factor.

Table 7-7. Factor matrix using a four factor structure

Item	Factor			
	1	2	3	4
Q1_agreed objectives for the session	.567			
Q2_ensured the trainee knew the role and name of each member of the endoscopy team	.440			
Q3_agreed and applied the ground rules	.655			-.330
Q4_questioned the trainee	.678			
Q5_provided explanations and descriptions	.600	-.346		
Q6_used a mixture suggestions, prompts, solutions and instructions	.644	-.400		
Q7_checked the trainee had understood instructions and advice	.610			
Q8_used an appropriate quantity of dialogue	.628	-.343		
Q9_asked the trainee to show where he or she was struggling	.500			
Q10_gave specific skills teaching	.623			
Q11_ did not overburden the trainee with too many tasks	.514		.541	.344
Q14_allowed the trainee reasonable time to carry out a procedure	.609		.364	
Q15_ always ensured the patient was comfortable and safe	.493	.397		
Q16_encouraged the trainee to communicate appropriately with the patient	.524	.573		.346
Q17_helped the trainee to assess if objectives for the session had been achieved	.638	.344		
Q18_reinforced positive aspects of the trainee's performance	.693			

Q19_identified aspects for the trainee to develop and improve	.606			

To get a better fit for the data the axis is rotated to give the pattern matrix (Table 7-8). This loaded seven items onto factor one, three items onto factor two, two items onto factor three, and six items onto factor four. All the items in the first factor appeared to match to the concept of dialogue. The second factor contained two items that referred to the patient and a third item that asked whether the trainer had helped the trainee to assess whether objectives had been achieved (Q17). This last item also loaded onto factor four with a higher loading and it was more theoretically in keeping with the other items in this factor; it was therefore allocated to factor four. The items in the third factor referred to the pace of the session and the items in the final factor focused on framing the learning.

Table 7-8. Pattern matrix of four factor extraction following rotation

Item	**Factor**			
	1	**2**	**3**	**4**
Q1_agreed objectives for the session				-.591
Q2_ensured the trainee knew the role and name of each member of the endoscopy team				-.365
Q3_agreed and applied the ground rules				-.806
Q4_questioned the trainee	.659			
Q5_provided explanations and descriptions	.768			
Q6_used a mixture suggestions, prompts, solutions and instructions	.828			
Q7_checked the trainee had understood instructions and advice	.475			
Q8_used an appropriate quantity of dialogue	.717			
Q9_asked the trainee to show where he or she was struggling	.407			
Q10_gave specific skills teaching	.533			

Q11_ did not overburden the trainee with too many tasks			.746	
Q14_allowed the trainee reasonable time to carry out a procedure			.485	
Q15_ always ensured the patient was comfortable and safe		.645		
Q16_encouraged the trainee to communicate appropriately with the patient		.930		
Q17_helped the trainee to assess if objectives for the session had been achieved		.371		-.499
Q18_reinforced positive aspects of the trainee's performance				-.534
Q19_identified aspects for the trainee to develop and improve				-.454

Regarding the statistical strength of these factors in terms of the factor loadings, Tabbachnick and Fiddell (2001) (cited in (Costello and Osbourne 2005)) recommend that all items should have loadings greater than 0.32; my data met this criteria. Stevens (2002 cited in Field 2009 pg 644) recommends that for a sample size of 200 a loading of greater than 0.364 is significant. My sample size was 166 and the lowest loading was .365 which is just above significance for a sample size of 200 (although our sample was slightly smaller than this).

A disadvantage of the above model was that two of the factors (factors 2 and 3) only contained two items each. Factors with less than three items are generally seen as weak and unstable (Costello and Osbourne 2005) meaning that with a greater sample size these factors may not be reproduced. I therefore re-ran the above factor analysis but removed these factors by specifying in SPSS that I only wanted to extract two factors rather than retain factors with eigenvalues over one. The factor matrix and pattern matrix are shown in Table 7-9 and Table 7-10.

Table 7-9. Factor matrix for two factor structure

Item	Factor	
	1	2
Q1_agreed objectives for the session	.560	
Q2_ensured the trainee knew the role and name of each member of the endoscopy team	.443	
Q3_agreed and applied the ground rules	.636	
Q4_questioned the trainee	.684	-.301
Q5_provided explanations and descriptions	.604	-.373
Q6_used a mixture suggestions, prompts, solutions and instructions	.647	-.414
Q7_checked the trainee had understood instructions and advice	.610	
Q8_used an appropriate quantity of dialogue	.627	-.325
Q9_asked the trainee to show where he or she was struggling	.506	
Q10_gave specific skills teaching	.627	
Q11_ did not overburden the trainee with too many tasks	.480	
Q14_allowed the trainee reasonable time to carry out a procedure	.598	
Q15_ always ensured the patient was comfortable and safe	.488	.373
Q16_encouraged the trainee to communicate appropriately with the patient	.493	.414
Q17_helped the trainee to assess if objectives for the session had been achieved	.632	.330
Q18_reinforced positive aspects of the trainee's performance	.696	
Q19_identified aspects for the trainee to develop and improve	.606	

Table 7-10. Pattern matrix for two factor extraction following rotation

Items	Factor	
	1	2
Q1_agreed objectives for the session	.309	.360
Q2_ensured the trainee knew the role and name of each member of the endoscopy team		.439
Q3_agreed and applied the ground rules		.481
Q4_questioned the trainee	.751	
Q5_provided explanations and descriptions	.754	
Q6_used a mixture suggestions, prompts, solutions and instructions	.820	
Q7_checked the trainee had understood instructions and advice	.492	
Q8_used an appropriate quantity of dialogue	.730	
Q9_asked the trainee to show where he or she was struggling	.444	
Q10_gave specific skills teaching	.602	
Q11_ did not overburden the trainee with too many tasks	.336	
Q14_allowed the trainee reasonable time to carry out a procedure	.344	.368
Q15_ always ensured the patient was comfortable and safe		.600
Q16_encouraged the trainee to communicate appropriately with the patient		.644
Q17_helped the trainee to assess if objectives for the session had been achieved		.619
Q18_reinforced positive aspects of the trainee's performance	.375	.459

Q19_identified aspects for the trainee to develop and improve	.385	.330

Using this two factor solution in the factor matrix all items loaded most strongly onto factor one but in the pattern matrix items did load onto both factors; however there were several problems with this new factor solution. Even with the pattern matrix four items loaded onto both factors; this is referred to as cross-loading and can be a sign that the items are poorly written or a flawed a priori factor structure(Costello and Osbourne 2005). Of these two possible reasons the latter is likely to be more responsible as the items have been extensively considered both through the cognitive interviewing and the Delphi process. Also the two factors were predetermined because of small numbers of items in the four factor solution not because of any theoretical reason. Several of the items also had low factor loadings. Additionally when items contained within the same factor were examined there did not appear to be obvious similar themes in terms of their content. Cronbach's alpha was then calculated for each domain, initially using the four-factor model. Cronbach's alpha for factor one, two, three and four were .865, .748, .664 and .807 respectively. All item-corrected domain correlations were above the accepted cut off of 0.3. Cronbach's alpha if item deleted could not be calculated for factor two and three as there were only two items in each domain (Table 7-11, Table 7-12, Table 7-13 and Table 7-14). For factor one if any item was deleted from that domain then Cronbach's alpha fell suggesting they all contributed to the internal consistency of that domain. In factor four Cronbach's alpha was actually marginally higher if Q2 (ensured the trainee knew the role and name of each member of the endoscopy team) was removed; this was perhaps not surprising as it also had a lower factor loading.

Table 7-11. Item-domain statistics for factor 1 using the four factor structure

Item	Corrected Item-domain Correlation	Cronbach's alpha if Item Deleted
Q4_questioned the trainee	.693	.838
Q5_provided explanations and descriptions	.675	.840

Q6_used a mixture suggestions, prompts, solutions and instructions	.735	.832
Q7_checked the trainee had understood instructions and advice	.546	.858
Q8_used an appropriate quantity of dialogue	.681	.839
Q9_asked the trainee to show where he or she was struggling	.503	.863
Q10_gave specific skills teaching	.627	.847

Table 7-12. . Item-domain statistics for factor 2 using the four factor structure

Item	Corrected Item-domain Correlation	Cronbach's alpha if Item Deleted
Q15_ always ensured the patient was comfortable and safe	.604	-
Q16_encouraged the trainee to communicate appropriately with the patient	.604	-

Table 7-13. Item-domain statistics for factor 3 using the four factor structure

Item	Corrected Item-Domain Correlation	Cronbach's alpha if Item Deleted
Q11_ did not overburden the trainee with too many tasks	.522	-
Q14_allowed the trainee reasonable time to carry out a procedure	.522	-

Table 7-14. . Item-domain statistics for factor 4 using the four factor structure

Item	Corrected Item-domain Correlation	Cronbach's alpha if Item Deleted
Q1_agreed objectives for the session	.534	.785

Q2_ensured the trainee knew the role and name of each member of the endoscopy team	.422	.809
Q3_agreed and applied the ground rules	.650	.760
Q17_helped the trainee to assess if objectives for the session had been achieved	.620	.765
Q18_reinforced positive aspects of the trainee's performance	.642	.761
Q19_identified aspects for the trainee to develop and improve	.549	.783

Using the two factor model the Cronbach's alpha for factors one and two were .868 and .826 respectively. All item-domain correlations were greater than 0.3 (Table 7-15 and Table 7-16). Only if Q11 was deleted from the factor 1 group did Cronbach's alpha increase, again this was an item that in the pattern loading had a low factor loading. Although for both the four and two factor models the Cronbach's alpha was lower than for the total this is likely to be due to fewer items in the domains than for the total and this reduces alpha. Cronbach's alpha was still greater than 0.8 for the domains containing more than two items and for those in the four factor model that only contained two items (factor 2 and 3) was still .748 and .664 which I felt was acceptable given the very small size of the domain.

Table 7-15. Item-domain statistics for factor 1 using the two factor structure

Item	Corrected Item-Domain Correlation	Cronbach's alpha if Item Deleted
Q4_questioned the trainee	.700	.845
Q5_provided explanations and descriptions	.667	.848
Q6_used a mixture suggestions, prompts, solutions and instructions	.689	.846
Q7_checked the trainee had understood instructions and advice	.582	.856
Q8_used an appropriate quantity of dialogue	.706	.843

Q9_asked the trainee to show where he or she was struggling	.530	.860
Q10_gave specific skills teaching	.622	.852
Q11_ did not overburden the trainee with too many tasks	.394	.873
Q19_identified aspects for the trainee to develop and improve	.533	.860

Table 7-16. Item-domain statistics for factor 2 using the two factor structure

Item	Corrected Item-Domain Correlation	Cronbach's alpha if Item Deleted
Q1_agreed objectives for the session	.517	.810
Q2_ensured the trainee knew the role and name of each member of the endoscopy team	.467	.816
Q3_agreed and applied the ground rules	.605	.798
Q14_allowed the trainee reasonable time to carry out a procedure	.500	.812
Q15_ always ensured the patient was comfortable and safe	.550	.805
Q16_encouraged the trainee to communicate appropriately with the patient	.542	.809
Q17_helped the trainee to assess if objectives for the session had been achieved	.632	.794
Q18_reinforced positive aspects of the trainee's performance	.594	.799

I have demonstrated that overall the tool has good internal consistency and that all attributes appear to be measuring a similar construct, this is demonstrated by the high overall Cronbach's alpha and also that the scree plot has suggested that there was one principal factor. Given these statistics it could therefore be argued that all the items

represent an overall construct of good endoscopy training. However what is unclear is whether within this construct of 'good endoscopy training' there are separate constructs that contribute to this overall construct. Using factor analysis appears to have produced four factors with content that makes sense and have reasonable consistency within each domain. However two of these domains were very small which is suggestive of an unstable factor structure. One solution would be to develop new items to include in factors two and three, however, this would not accord with the view of the Delphi panel that there should be as few items as possible. Alternatively, restricting the analysis to extract only two factors led to statistical flaws and the items in each factor were less theoretically coherent. As all the items correlated in the following generalisability analysis case scores were used rather than item or domain scores.

7.4.4 Generalisability

Using mean scores, in order to include missing data, the variance components for the first generalisability analysis are shown in Table 7-17 with 44% of the variance in scores explained by true differences in trainer ability; 34% of variance due to reviewer stringency and 22% due to reviewers marking particular trainers in particular ways.

Table 7-17. . Variance components expressed as numbers and percentages for DOTS peer data using main expected sources of variance

Component	Sum of squares	Degrees of freedom	Mean squares	Estimate	Percentage variance
σ^2_p (Trainer)	14.4	37	0.39	0.11	44%
σ^2_r (Reviewer)	13.5	52	0.26	0.09	34%
σ^2_{tp} (Trainer*reviewer)	6.3	92	0.07	0.06	22%
Error	0	0	0	0.00	0%

The G co-efficients calculated using the above variance components for differing numbers of reviewers are shown in Table 7-18; three reviewers would be needed to gain a reliability of 0.7. Table 7-18 also shows the SEM and the 95% confidence intervals that would be expected with differing numbers of reviewers. Thus for three reviewers the total score might vary by nearly ten points in either direction.

Table 7-18. . G co-efficients using differing numbers of reviewers considering trainer, reviewer and reviewer: peer interaction as the sources of variance

Number of reviewers	G co-efficient	SEM	95% CI
1	0.44	6.74	13.21
2	0.61	5.63	11.03
3	0.70	4.93	9.66
4	0.76	4.41	8.65
5	0.80	4.03	7.90
6	0.82	3.82	7.49
7	0.84	3.60	7.06
8	0.86	3.37	6.61
9	0.87	3.25	6.37
10	0.89	2.99	5.86

The results of the generalisability analysis using all possible sources of variance are shown in Table 7-19. There were five sources of variability in total (Table 7-19) including some surprising sources of variance. Our previous hypothesised sources of variation remained but accounted for much less of the variance than had previously been assumed. Further sources of variance that were identified were variation in scores due to course and due to an interaction between the reviewer marking trainers more or less stringently due to the trainer's speciality.

Table 7-19. Variance components for the DOTS when completed by peers accounting for all possible sources of variance

Component	Sum of Squares	Degrees of freedom	Mean squares	Estimate of variance	Percentage variance
σ^2_p (Trainer)	8.29	30	0.28	0.074	29%
σ^2_r (Reviewer)	12.03	48	0.25	0.028	11%
σ^2_{course} (Course)	0	0	0	0.057	22%
σ^2_{tp} (Trainer*reviewer)	4.57	70	.065	0.054	21%
$\sigma^2_{r\ spec\ trainer}$ (Reviewer * Specialty_trainer)	0	0	0	0.042	16%
Error	0	0	0	0	

As can be seen from Table 7-19 however, degrees of freedom were not reported for course or reviewer* speciality of trainer interaction. Degrees of freedom are not reported for course as all other variables are nested within the course and there was no crossover of individuals between courses; this means it is not possible to look at variance for course. Speciality of trainer is also nested within trainer and therefore is already accounted for by the trainer*reviewer interaction within each course where this is the case SPSS does not report degrees of freedom or sum of squares. Although Table 7-3 and Table 7-4 suggest that there were differences in scores between courses, given the format in which the data has been collected, it suggests that the first simpler analysis better explains the variance in scores.

When the first analysis was re-run using the total score rather than mean the results were comparable to those shown above; there were slight changes in the absolute numbers of the variance components and their percentages but not to the order in terms of percentage contribution to the variance (G co-efficient for one reviewer 0.50 using total scores).

7.5 Discussion

Although I expected each trainer to receive the same number of evaluations this differed markedly. This was because I had directed peers to evaluate the trainer only if they saw the case in its entirety, because this is a very busy day on the course sometimes a debrief of the last case occurs at the same time as the next case starts therefore the trainer and faculty member involved in the debriefing miss the start of the next case and therefore were unable to evaluate that aspect of training. This may also account for some of the missing data as these elements of the case may not have occurred in an observable area and therefore the peers were not able to comment. For instance item 2

The trainer ensured the trainee knew the name and role of each member of the endoscopy team before a training encounter so that the trainee was supported

was the item that was missed the most frequently, this may have been because peers did not see this occurring but were concerned that the trainer may have done this when they were not observing and therefore were unwilling to disagree with the statement. An alternative explanation may have been that the peers did not

understand this item. Before commenting further on the missing data it would also be worthwhile looking at whether similar data is missing when trainees complete the tool and when it is completed by trainers as a self-assessment exercise. The fact that not all reviewers reviewed all trainers is a limitation of this study and demonstrates a disadvantage of using the 'Training the Trainer' courses as a method by which to trial the tool.

This is not the only disadvantage of using the TCT course to validate the tool. I have already discussed the concept of ecological validity which is lacking on the TCT course although all trainers that participated in the TCT course were also trainers within their own department. Also on the TCT courses often one of the other course attenders acts as a 'trainee' in that they are taught colonoscopy by one of the other trainers. This will clearly alter the training interaction between 'trainee' and trainer and this may alter how the trainer trains as they are already teaching someone who is competent at endoscopy and will have affected the validity of the test situation. The DOTS however was still able to discriminate between different trainers and appeared reliable in doing so, regardless of whom the trainer was training. It would however be important to ensure the tool was still reliable when used by a peer evaluating a trainer who was training a novice trainee as this may alter the training dynamic and hence the reliability of the tool.

It is also important to note that the above data pertained to colonoscopy training only and therefore caution must be taken in generalizing in to all endoscopy training. Trainees do however spend much longer learning colonoscopy than upper GI endoscopy as it is more technically demanding. There are many more colonoscopy 'Training the Trainer' courses. One upper GI endoscopy course did occur during the time period this research was undertaken however the course leader did not respond to email invitation to take part.

A wide range of scores (31 to 85) was given to trainers. Item analysis showed that for 14 of the 17 items all points on the scale had been used. However the top end of the scale was used preferentially for all items; this was demonstrated by the test for skewness which was negative for all items. It is not uncommon to find this negative skew on rating scales (Beckman, Ghosh et al. 2004). Examining the data at the individual reviewer level the majority of the reviewers awarded trainers a range of

scores. On initial inspection there did appear to be a couple of reviewers who awarded the same total score to more than one trainer however on closer inspection there still remained variety in item scores. Trainers aggregated scores followed a normal distribution although this was not centred over the midpoint of the possible range of scores (Figure 7-1)This suggests that the DOTS tool enables reviewers to discriminate between trainers. The mean and standard deviations differed when the data was examined for the different courses; for some of the courses the confidence intervals did not overlap; for instance course 100 and 400. When a one-way ANOVA was performed there were significant differences between groups. Reasons why this might be are explored in section 9.3.2

7.5.1 Internal structure

Trialling the toolkit for peers enabled evidence to be obtained for the internal structure of the tool. The toolkit demonstrated high internal consistency with a high Cronbach's alpha at 0.895. This was somewhat expected given the large number of items however all items contributed to the high alpha; this is demonstrated by the fact that if any one item was deleted from the tool then the size of alpha decreased. All items demonstrated moderate item-corrected total correlations of greater than 0.4. These statistics in combination suggest that all items contribute to a shared variance and are measuring aspects of the intended construct of being a good trainer. As discussed in Chapter 6 however it could be argued that this only demonstrates item equivalence rather than homogeneity in that these statistical tests measure the amount of uniqueness compared to shared variance contained in each item. In order to examine further for homogeneity factor analysis was performed.

The scree plot was strongly supportive of item homogeneity as it suggested that much of the shared variance could be explained by one factor which could be argued to represent teaching proficiency. This one factor accounted for 38% of the variance. When I extracted items with eigenvalues greater than one this resulted in a four- factor structure which accounted for 61% of the variance. When I examined the items within each of these four factors there does appear to be some similarity between items grouped within the same factor. For instance the seven items within factor one all pertain to elements of dialogue. Two of the items in the second factor refer to the patient and ensuring the training episode remains focused on the patient. The third item in this factor however, refers to helping the trainee assess if the objectives for the

session have been met does not fit with this 'patient focused' interpretation. The items in the third factor both refer to the pace of the session and the items in the fourth factor are about the setting of the session. An alternative explanation of the factors though is that the items were clustered in this manner because they represent the timing of the session i.e. the preparation, the teaching during the case and the teaching at the end. Alternatively the tool may have loaded onto these four factors in this way simply because of the order of the items on the tool as items loaded onto the factors sequentially down the tool. As I did not randomise the items the factor structure may be displaying an order effect. When these factors were examined further each factor had an acceptable level of Cronbach's alpha with all but factor three having an alpha greater than 0.7 and item-total correlations of greater than 0.3; this suggests that shared variance did exist between items that loaded onto the same factor. In terms of the stability of the factor structure itself, all items loaded onto the factors with loadings greater than 0.3 which is supportive of factor stability, and only one item cross-loaded onto two factors. Factor two and three however only contained two items each, this is suggestive that these factors are weaker and might not be found again on successive data.

To try and overcome this, the factor analysis was repeated but, in order to exclude these factors, I specified that only two factors be extracted. This obviously led to a two factor structure and although both of these factors had reasonably high Cronbach's alpha (0.868 and 0.826) several of the items loaded onto a factor with a loading of less than 0.3 and there were several factor crossloadings making this structure unstable. In addition to this, the grouping of the items into the two different factors did not appear to have any coherence of content. As this two-factor structure does not appear to be suitable an alternative solution to overcome the problem with the four factor model would be to try and add more items that were in keeping with the two smaller factors in the four factor model. However within the Delphi process the largest voice from the panel was that the tool should be as short as possible and therefore adding items would detract from this need. An alternative method to examine the factor structure further would be to collect more data and see if the same factor structure existed on a repeat sample as demonstrated by the Litzelman et al (1998) example discussed in Chapter 6. It is also important to note that in fact this is only a statistical interpretation of the factors and therefore any theoretical interpretation must be made very

cautiously (Field 2009). I will therefore discuss this in further detail in chapter 10 when I discuss how the results of the toolkit might be fed back to trainers.

7.5.2 Generalisability analysis

The initial generalisability analysis demonstrated that the largest variation in scores was caused by differences in trainers i.e. their differing abilities to train which was what I had intended the tool to measure; this accounted for 44% of the variance. Reviewer stringency also explained some of the variance as did the interaction between trainer and reviewer. From this I mean that differences in trainers aside from their true ability as trainers caused certain reviewers to give them different marks. This interaction resulting in variance is not unexpected and is seen in theoretical textbooks (Shavelson and Webb 1991). This final variance component also includes other sources of variance that we have not been able to define, i.e. unexplained errors of variance. Using this model it is possible to predict that we would need three reviewers to gain a reliability of 0.7 which Downing (2004) recommends for a formative tool. When I looked at all potential sources of variability I found that there were more components that explained variance than I had hypothesised however this model was of poor fit indicated by the negative variance components and failure of SPSS to report degrees of freedom. These issues were explored by deleting variables that produced a negative variance component and re-running the analysis. Only two variables remained (course and reviewer* speciality of trainer) and these were rejected when explored using a type III ANOVA as suggested by Crossley, Russell et al (2007). This can be explained by the variables being nested in course and reviewer* speciality of trainer being confounded by the trainer*reviewer interaction. This suggests that the simpler analysis best describes the sources of variance.

In trialling the tools on the TCT courses I have made a judgement that all other participants are valid peers. This presumption was based on the fact that all course attendees were trainers within their own base units but also they had attended day one of the course which had contained theory about the practice of teaching endoscopy. This teaching gave them the knowledge to appreciate the subtleties of teaching endoscopy. It does however mean on the second day of the course when the evaluations took place that the trainers were likely to put these newly taught skills into action and train as they had been taught to do so. It is also likely that the peer reviewers also judged the teaching session in reference to the theory they had learnt

the day before. This means that there is a risk of 'cultural reproduction' where pre-existing theories are perpetuated, however the toolkit was made from the opinions of many and could still pick up differences in technique but it may have falsely elevated the reliability as everyone is teaching in the 'same way' and may have introduced a source of bias.

7.6 Conclusion

Trialling the DOTS tool for peer evaluations on the TTT courses has given a useful insight into how peers use the tool and the internal structure of the tool. Not all the items were consistently answered; there are likely to be some items that were accidently missed (Acock 2005) but items like item 2 have been more systematically missed. This might be due to the nature of courses rather than due to how the peers completed the tool and this demonstrates a disadvantage of not trialling the tool within its naturalistic setting.

There is evidence that the items demonstrate homogeneity and certainly contain a shared amount of variance but there is also an argument that domains do exist within the data. There are however issues with the stability of these domains as discussed above, in terms of the four factor structure a potential source of instability could be that two of the factors were very small; this will be discussed further in Chapter 9.

The tool appeared to show reasonable reliability; with three reviewers required to gain a reliability of greater than 0.7.

In the next chapter I will discuss trialling the DOTS and LETS for use by trainees and trainers as a self-assessment exercise within local units; a setting with greater ecological validity. In chapter 9 I will then reflect on the results of both trials in combination.

Chapter 8. Trialling the Toolkit within Local Units

In the previous chapter I discussed the evidence for the reliability and internal structure of the DOTS tool when it was used by peers to evaluate trainers. The DOTS tool was also intended to be used by trainees and trainers, as a self-evaluation tool, and therefore I needed to evaluate the tool for these purposes too. Additionally I wanted to gain an assessment of the internal structure and the reliability of the LETS. In this chapter I will discuss how I trialled the tools within local units. As well as examining the internal Structure of the toolkit I still needed to provide evidence for its relationship to other variables. Trialling both the LETS and DOTS in the same sample population allowed for a comparison between the two tools, however as neither of these are validated I also needed to compare the tools to a previously validated tool, this is also discussed within this Chapter.

8.1 Study Design

In trialling the DOTS and the LETS it was important to do so in local endoscopy units under the conditions in which training would actually occur. The reasons for doing this is that clearly it would be difficult to examine the LETS under any other setting as it aims to evaluate training over the longer term and therefore it would be impossible to test it on a course where no such extended training relationship occurs. I also wanted to test the DOTS under the same situation as the LETS in order to be able to correlate the two tools to examine the extent that they measure the same construct; this required trainers to be using the tools with the same trainees in order to examine that statistical relationship. In addition to these reasons I felt that testing the tools within practicing endoscopy units would create greater ecological validity than just trialling the tools on courses. As discussed ecological validity refers to the process of trying to match the experimental conditions of a test to those most closely matching real life(Cohen, Manion et al. 2007).

Testing the toolkit within these settings led to issues with the design of the study; because it was the 'real world' it was not possible to optimise the setting in order to gain the best design to examine the reliability and internal structure of the tools. The organisation of endoscopy training varies from NHS Trust to NHS Trust in terms of how trainees are allocated to trainers or certain training lists. For instance, in some Trusts

trainees are not allocated to certain trainers and attend the lists that fit in with the rest of their timetabled commitments; this means for the trainer that they may have a different trainee attending each of their lists. For the majority of trainers however, a trainee is allocated to a trainer and attends their endoscopy list every week. This means that for the majority of endoscopy trainers they only have one or at the most two trainees at any one time.

This therefore makes it more difficult to look at certain aspects of reliability; for instance it is more difficult to look at the reliability of ratings given by different trainees (inter-rater reliability) in terms of both the DOTS and the LETS as a trainer may only be training one trainee at a time. It would be possible to consider test-retest reliability for the DOTS by asking a trainee to complete the evaluation on more than one occasion. For the LETS even trying to gain an estimate of the test-retest reliability would be difficult as it should be completed after an extended period of training, as a trainer may only have one trainee over the duration of a year it would be difficult to gain multiple LETS within the time constraints of this study. There were also difficulties in considering what reliability can be gathered from the self-evaluations for both tools, clearly there is only one self so it would not be possible to look at the inter-rater reliability. It would be possible to consider the test-retest reliability for the DOTS but this carries the same problems as the trainee evaluations for the LETS. Given these difficulties in data collection it was decided that classical test theory would be used to analyse the data, this is because there was too little crossover between the data and all trainees would be nested in trainers which made it unsuitable for the use of Generalisability theory. Classical test theory meant that it was still possible to gain an assessment of the reliability of the data collected but only one variable can be examined at one time. The variables it was possible to measure was test-retest reliability for the DOTS and inter-rater reliability between trainees and trainers for both the LETS and the DOTS.

8.1.1 Relationship to other variables

An advantage of testing the tools in local units was that it would be possible to compare the differences between trainees and self-evaluations and between the DOTS and LETS. This would contribute to the examination of the validity of the toolkit by considering the source of evidence referred to as the 'relationship to other variables'(Downing 2003). However as the DOTS and LETS form part of the same toolkit

and were developed together I felt a comparison between the two tools was inadequate to fully explore this category of validity evidence. In addition neither of the tools had been previously validated. To provide further evidence that these tools measured good clinical teaching it was necessary to compare the toolkit to a previously validated tool(Beckman, Cook et al. 2005). I opted to use the LETS for comparison pragmatically as it would be required to be completed less frequently and therefore, if asking respondents to complete a further tool as well, would create less of a respondent burden. I therefore required a validated tool that also looked at teaching behaviour over the duration of a rotation.

Comparison of the LETS with an already established instrument is a powerful assessment of convergent validity (Beckman, Lee et al. 2004). One of the options would have been to use the endoscopy trainer tool already in existence on the JETS website(JAG 2012), however as this has no published data regarding its validity I did not feel that this would provide strong enough evidence of validity. I initially considered the surgical tools discussed in Chapter 2 as I felt that surgical teaching was most similar to endoscopy teaching as it also involves teaching a complex practical skill. On review of the surgical evaluation tools some were not appropriate as they were for use after a single session(Hauge, Wanzek et al. 2001, Sarker, Vincent et al. 2005) or made no assessment of the tool's reliability (Cox and Swanson 2002, Claridge, Calland et al. 2003, Iwaszkiewicz, DaRosa et al. 2008). Two of the remaining tools (Downing, English et al. 1983, Maker, Curtis et al. 2004) considered surgical teaching behaviours which were not relevant to endoscopy; this was mainly because they specifically mentioned teaching in other physical environments in which an endoscopy trainer would not teach their trainee, for instance on wards or teaching conferences. One option would be to remove these items but then the current evidence for validity of that tool would not be valid. The final two tools (Risucci, Lutsky et al. 1992, Cohen, MacRae et al. 1996) could be used to compare to the LETS as both tools have evidence of reliability and all the items could apply to an endoscopy trainer. Cohen, McRae et al.'s (1996) tool can be used over a rotation and only contains four items so is very short and therefore if I asked trainees to complete this as well as the LETS it would only be a small increase in participant burden, however this tool has only ever been used with students in one institution and therefore its reliability pertains to this use. As the validity evidence of a tool only refers to the validity of the tool under the circumstances it has been tested (Downing 2003), I felt that in using the tool with trainees rather than students, in a

different institution and a different specialty I would be significantly altering the test circumstances and therefore it may be difficult to draw conclusions from its validity. I had a similar concern over the tool that had been designed by Risucci, Lutsky et al (1992) that, although it had been used with residents, it had only ever been used within one institution.

I therefore decided to look more widely at clinical teaching tools rather than just tools that focused on surgical teachers. Many of the tools mentioned in the Beckman, Cook et al (2005) and the Fluit, Bohuis et al (2010) reviews were not appropriate; some of these looked at multiple teachers rather than the individual teacher (Roff, McAleer et al. 2005) or also considered environmental factors such as learning resources (James and Osborne 1999) which had been excluded from the toolkit as they were outwith the trainer's control. Other tools considered areas that were not relevant to the endoscopy trainer or had only been trialled in one institution. The two tools that were identified as potentially suitable were the Clinical Teaching Effectiveness Inventory (CTEI)(Copeland and Hewson 2000) or the SFDP26 developed by Litzelman, Stratos et al (1998). Both of these tools contained items that were generic enough to apply to endoscopy trainers and were not environmentally specific. There was no evidence that either tool had been used with endoscopy trainers but both tools had been trialled across specialities and used in more than one institution. The CTEI was chosen as it is shorter in length (15 items vs. 26 items) and, as this method requires participants to complete both the LETS and the established instrument, brevity was felt to be an important factor to try and reduce participant fatigue. The CTEI has been reported to show reasonable reliability; Copeland et al (2000) report a G co-efficient of 0.74 with just one rater.

8.2 Methodology

8.2.1 Data collection

I initially aimed for both elements of the toolkit to be completed online, therefore both the DOTS and LETS with the CTEI were uploaded to Survey Monkey (Surveymonkey 2008) (Appendix 7 and 8).

The two tools were trialled within the Northern Deanery using the Northern Region Endoscopy Group (NREG) to make contact with endoscopy trainers. An invitational email was sent via the Chairman of NREG which detailed the purpose of the project and

contained my email address to respond to should they wish to participate. Once a respondent replied they were sent more information about the study along with the URL for both the DOTS and the LETS. Each trainer was also sent a trainer code which acted as a unique identifier for that trainer. The trainer was asked to complete the DOTS and LETS with each trainee they currently had contact with, as a self-evaluation, and also send the login details for the two tools to each trainee for them to complete along with an information sheet for the trainees. The trainers were asked to disseminate information to their trainees in this way to ensure that the trainer had consented to be evaluated by their trainee; this was confirmed by the fact that the trainee would use the trainer's code. It was emphasised that as long as the trainer had been training the trainee for more than two months then a LETS could be completed at any time and that the DOTS should be completed by trainer and trainee after the same list. It was also requested that the DOTS be completed on two occasions by both the trainer and trainee in order to gain an assessment of test-retest reliability.

There were separate URLs for the DOTS and the LETS but both the trainer and trainee used the same URLs for each. Each tool initially asked for the trainer's code and then asked whether the respondent was the trainer or trainee; this question then contained a skip link that took the respondent to the correct page; the tools for the trainer and trainee only differed in that, for the trainer, the stem was in the first person.

Demographics for the respondent were collected. If the respondent was the trainer then they were asked for their specialty (gastroenterology, surgery, nurse or other) and whether they had attended a 'Training the Trainer' course. The trainee was asked for their name (which was later anonymised), their specialty and the numbers of procedures they had performed. For the DOTS both the trainer and trainee were asked to give the date of the list in order to match up their responses and the type of list; endoscopy, colonoscopy, ERCP or mixed list. The LETS also contained the CTEI which respondents were asked to also complete.

Due to a limited response electronically, paper copies were also posted to trainers. Each trainer received a pack which contained a letter explaining the study and containing their unique trainer code, two copies of the DOTS for the trainer to complete and two copies for the trainee to complete; each paper copy of the tool also had a prepaid envelope attached to send the completed tools back to myself. The pack also contained a letter that could be returned if the individual was not interested in

participating or did not currently have a trainee attached to them. The pack only contained the DOTS as this already meant quite a lot of paper and I did not want to overburden them; once the trainer had returned a DOTS they were sent an email asking them to also complete a LETS which contained the URL and a reminder of their trainer code.

8.2.2 Analysis

Trainees were given a unique identifier in order to anonymise the data (trainers already had a unique identifier). The data was then entered into a database on SPSS version 14 (SPSS 2005). Demographics of respondents were recorded and responses were analysed.

The DOTS was examined initially. The mean, median and range of scores were initially explored and the data examined for normality for all the evaluations initially. Normality was examined using Q-Q plot and calculating z-scores for skew and kurtosis (Field 23009 pg 139). The process was then repeated but looking at trainee and trainer evaluations separately. Individual respondents were examined to look for halo effect. Halo effect is the act of giving an individual the same score for every item based on the respondent's general impression of the individual rather than a specific answer for every question (Streiner and Norman 2008) p121). Scale frequency for each item was also reviewed to examine how the scale had been used by the respondents. If all respondents had given the same score for an item then this item is said to have poor discriminating power (Streiner and Norman 2008)

As in chapter 7, the internal structure of the tool was examined. Item-total correlations were calculated with the item in question deleted (item-corrected total correlation). Cronbach's alpha was calculated for the tool overall, the alpha if the item deleted for each item was also calculated. This was performed using all the DOTS data and then the data was re-analysed considering only the trainers' self-evaluation scores and then the trainee evaluations.

The inter-rater reliability between the trainee evaluation and the trainer's self-evaluation was calculated using the Pearson r correlation for parametric tests and was then repeated using the Spearman Rho for non-parametric tests. This is because as discussed in section 5.1.1.1.2 Likert scares are technically ordinal data and therefore should be examined using non-parametric tests however if the data is not skewed it is

acceptable to analyse the data as interval data. Both tests were performed as two-tailed tests.

Using all the DOTS evaluations (including both the trainer and trainee evaluations) the test-retest reliability was calculated again using both the Pearson correlation and Spearman's test, both as two tailed tests. The median time between evaluations was also calculated as correlations would expected to be lower the longer the interval between evaluations.

The LETS was then examined in the same way as the DOTS. The mean, median and range of scores were calculated and tests for normality were performed. The internal structure of the tool was also examined in the same manner as the DOTS described above. Inter-rater reliability between trainee and trainer scores was also calculated in the same way as for the DOTS

The association between the scores for the LETS and the DOTS, and the LETS and CTEI was examined for the same trainers using Pearson r and Spearman Rho correlations.

Free text comments were counted and reviewed. The comments were then thematically analysed to look for common themes.

8.3 Results

8.3.1 Demographics

It is difficult to estimate how many trainers were emailed about the study by the chairman of NREG as many on the list may not have been trainers and may or may not have disseminated to other trainers within their trust. Paper copies of the DOTS tool were posted to 60 surgeons, gastroenterologists, and nurse endoscopists around the region, although not all of these may currently practice as endoscopy trainers. Ten respondents returned the respondent slip to state that they did not currently have a trainee attached to them or were not currently practicing endoscopy.

Eleven trainers participated in the study and I received at least one completed evaluation tool from them and their trainee. Of these 11 trainers two were nurse endoscopists and the rest were consultant gastroenterologists. All but one participant had attended a 'Training the trainer' course.

8.3.2 DOTS

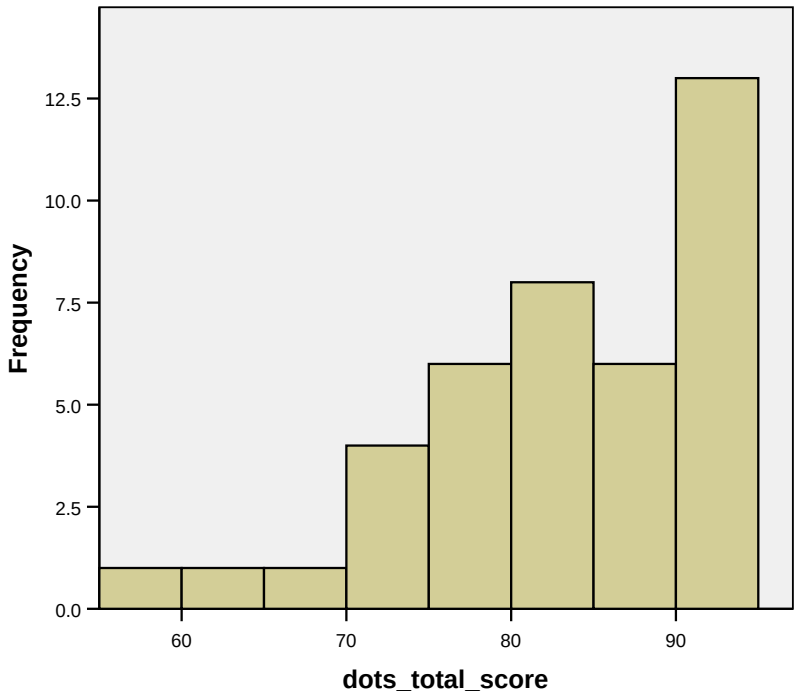
Forty DOTS evaluations were performed in total, one trainee evaluation contained a missing data point and therefore was not included in the initial analysis. In order to enable this evaluation to be included a mean score was calculated for the missing data point using the scores that that trainee had given the trainer for all the other items and the analysis re-run. The following results represent this analysis.

The spread of scores for the DOTS is seen in Table 8-1 with a range from 58 to 95 (the maximum score possible compared to when the tool was trialled by peers on the training courses where the maximum score was 85 as two items were omitted) and a median score of 82. A histogram of the spread of scores is also shown Figure 8-1 . This demonstrates that there appears to be a reasonable spread of scores but that scores are negatively skewed as confirmed by the negative skewness score in Table 8-1.

Table 8-1. Descriptive statistics for the DOTS when used in local units

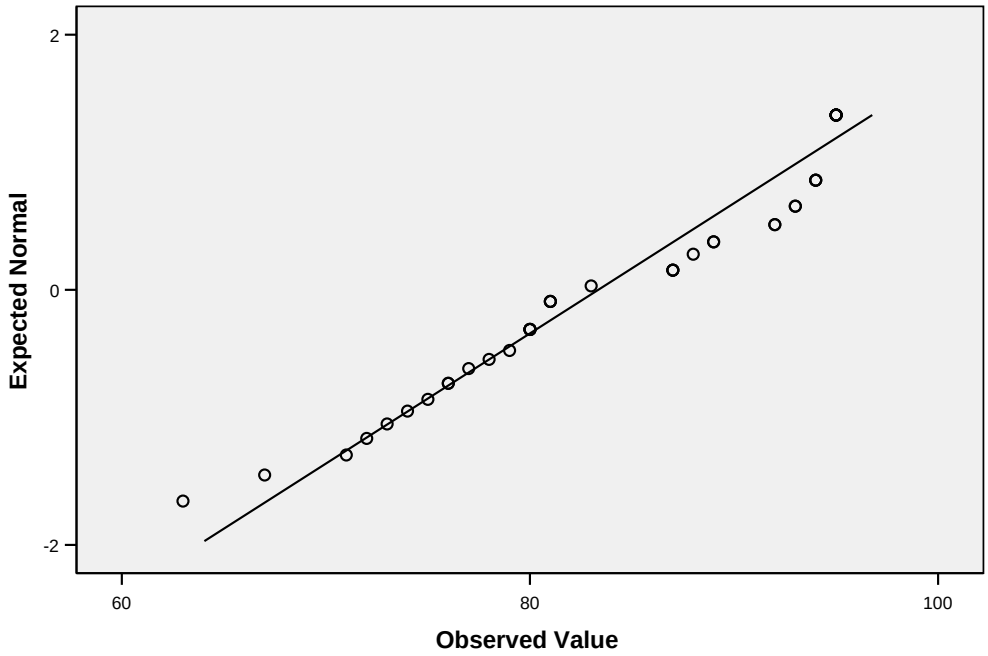
Descriptor	**Statistic**
Mean (standard error)	83.35 (1.549)
Median	82.00
Variance	95.977
Standard deviation	9.797
Minimum	58
Maximum	95
Range	37
Interquartile range	17
Skewness	-0.563
Kurtosis	-0.292

Figure 8-1. Histogram of frequency of DOTS scores when used in local units



Given the shape of the histogram in Figure 8-1 I wanted to explore the data for normality as this would affect the choice of statistical tests subsequently used to further analyse the data. Normality is also important as it suggests that the tool has been trialled across, and is able to detect a spectrum of performance. The z-score for skewness was calculated as 1.51 and for kurtosis was 0.40 both of these are not significant at the $p<.05$ level suggesting that they were not significantly different from a normal distribution given the number of responses. A Q-Q plot (Figure 8-2) shows reasonable but not perfect fit for normality. For this reason both parametric and non-parametric tests for correlation were carried out.

Figure 8-2. Normal Q-Q plot for the DOTS when used in local units

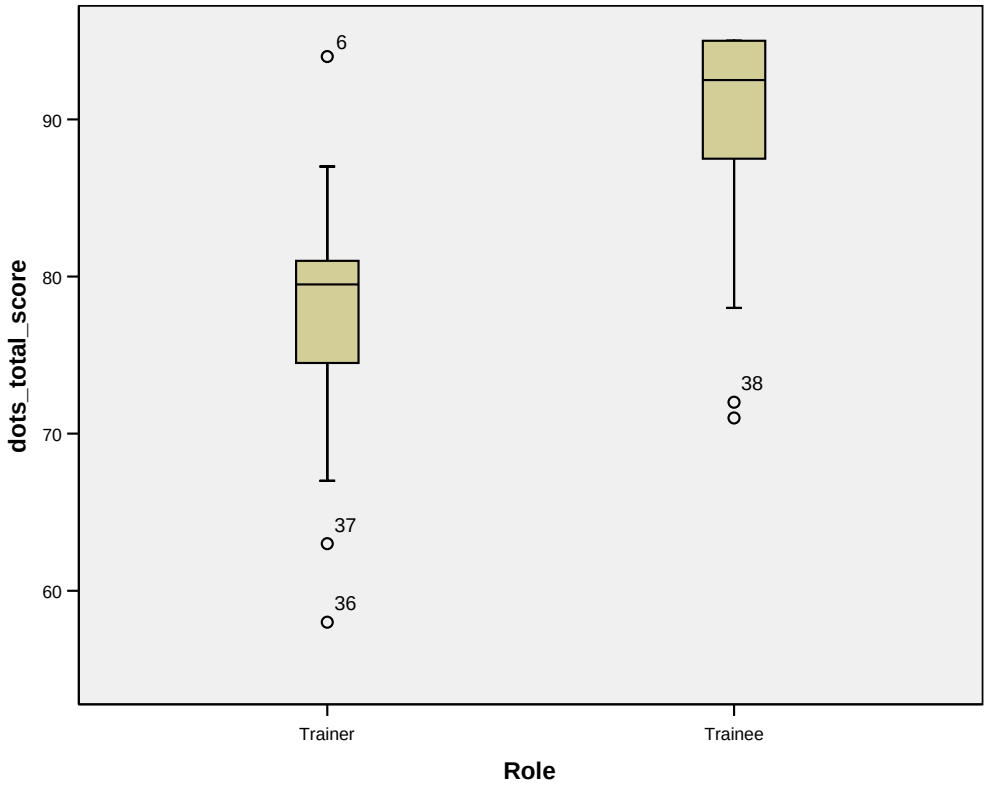


The DOTS data was further explored considering the trainee and trainer scores separately. There were 20 trainer evaluations and 20 trainee evaluations. The range, median and mean scores are shown in Table 8-2. As can be seen from this table the trainees tended to give higher scores compared to trainers self-evaluation scores and used a narrower range of scores; this is also clearly visible in the boxplot in Figure 8-3.

Table 8-2. Descriptive statistics of trainee and trainer scores for the DOTS when used in local units

Descriptor	Trainer data	Trainee data
Mean (standard error)	77.60 (1.839)	89.1 (1.729)
Median	79.50	92.50
Variance	67.621	59.779
Standard deviation	8.223	7.732
Minimum	58	71
Maximum	94	95
Range	36	24
Interquartile range	7	8
Skewness	-0.590	-1.446
Kurtosis	1.131	0.987

Figure 8-3. Box plot comparing trainee and trainer evaluations for the DOTS when used in local units



The trainer and trainee evaluations were also examined for normality; z scores for skewness and kurtosis were calculate as 2.82 and 0.99 respectively for trainees and 1.15 and 1.40 for trainers; this is above the level of significance 1.96 for the $p<0.05$ level (Field 2009) for trainee data suggesting this is not normally distributed. This adds further argument to analysing the data using both parametric and non-parametric tests.

Given the shape of the histogram in Figure 8-1 it was clear that respondents had tended to score trainers highly. When the items were reviewed individually there was no item that had been scored the same by all respondents but there was a tendency to use the top half of the scale. For four items only the top half of the scale was used (Q1, Q6, Q7 and Q8) and for a further four items only neutral and the upper half of the scale was used(Q4, Q5,Q14 and Q15). When this was considered for trainees and trainers separately, both groups used the top two options for only a quarter of the items but trainees were much less likely to use the lower points (disagree or strongly disagree) compared to trainers; trainers used the bottom half of the scale for ten of the 19 items compared to only two items by trainees.

In considering respondents individual answers, one trainer had agreed with all the items when completing their self-evaluation and four trainees had strongly agreed with all the items; this meant that 36% of trainees only used the top point of the scale on one evaluation. All these respondents however had completed two evaluations and on the other evaluation had given a more varied response, although admittedly this tended to be a mixture of agree and strongly agree so remained in the same half of the scale.

8.3.2.1 Internal structure of the DOTS

Cronbach's alpha considering trainee and trainer evaluations in combination was 0.945. The item-corrected total correlations and the Cronbach's alpha if item-deleted are shown in the first two columns of Table 8-3. All corrected item total correlations were greater than 0.3 suggesting adequate item total correlations. The Cronbach's alpha if just trainer evaluations were examined was 0.907 and for trainee evaluations was 0.934. Considering all the item-total correlations there appeared to be a mix of items that correlated poorly but this was not consistent between the two groups. One of the items that correlated poorly when evaluated as part of a trainer self-evaluation was item 2 (The trainer ensured the trainee knew the role and name of each member of the endoscopy team so that the trainee felt supported). In contrast this item correlated well with the total as part of a trainee evaluation. Item 14 (the trainer allowed the trainee reasonable time to carry out the procedure) correlated poorly with the total when evaluated by trainees. These items increased Cronbach's alpha if the item was deleted when considering total evaluations for the respective group of respondents.

Table 8-3. Internal structure statistics for the DOTS tool used in local units comparing all data, trainee evaluations and trainer evaluations

Item	All Evaluations (Cronbach's alpha 0.945)		Trainer Evaluations (Cronbach's alpha 0.907)		Trainee evaluations (Cronbach's alpha 0.937)	
	Corrected item-total correlation	Alpha if item deleted	Corrected item-total correlation	Alpha if item deleted	Corrected item-total correlation	Alpha if item deleted
Q1_agreed objectives for the session	.586	.944	.522	.904	.574	.932
Q2_ensured the trainee knew the	.486	.948	.039	.922	.673	.930

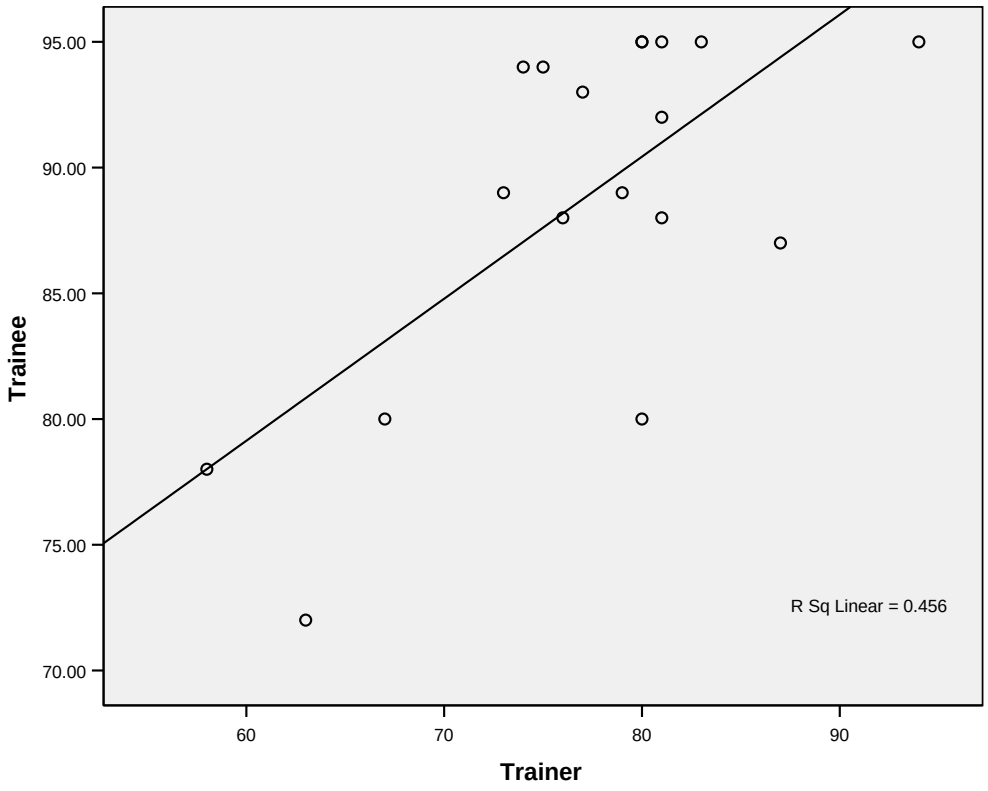
role and name of each member of the endoscopy team						
Q3_agreed and applied the ground rules	.640	.943	.535	.903	.797	.928
Q4_questioned the trainee	.670	.943	.513	.905	.526	.933
Q5_provided explanations and descriptions	.793	.942	.513	.905	.838	.929
Q6_used a mixture suggestions, prompts, solutions and instructions	.698	.942	.477	.905	.677	.930
Q7_checked the trainee had understood instructions and advice	.707	.942	.648	.901	.514	.933
Q8_used an appropriate quantity of dialogue	.698	.942	.677	.901	.323	.935
Q9_asked the trainee to show where he or she was struggling	.775	.941	.778	.869	.680	.930
Q10_gave specific skills teaching	.737	.941	.665	.899	.772	.928
Q11_ did not overburden the trainee with too many tasks	.592	.944	.139	.913	.835	.928
Q12_demonstrated a procedure	.801	.940	.800	.895	.705	.929
Q13_ intervened in a timely fashion	.702	.942	.533	.903	.793	.927
Q14_allowed the trainee reasonable time to carry out a procedure	.435	.946	.339	.907	.268	.936

Q15_ always ensured the patient was comfortable and safe	.702	.943	.555	.903	.835	.928
Q16_encouraged the trainee to communicate appropriately with the patient	.781	.941	.799	.894	.651	.931
Q17_helped the trainee to assess if objectives for the session had been achieved	.788	.940	.797	.895	.655	.930
DOTS_Q18_reinforced positive aspects of the trainee's performance	.712	.942	.791	.895	.484	.933
Q19_identified aspects for the trainee to develop and improve	.791	.940	.835	.894	.703	.929

8.3.2.2 Inter-rater reliability

The Pearson correlation of trainer and trainee evaluations was 0.676 which was significant at the 0.01 level. The Spearman rho correlation was 0.516 which was significant at the 0.05 level. This correlation is demonstrated graphically in Figure 8-4; although the correlation has reached significance there are not actually that many points on the line, rather most of the results are clustered at the top end of the score range. The significance may therefore have been artificially affected by the small number of lower scores by increasing the range.

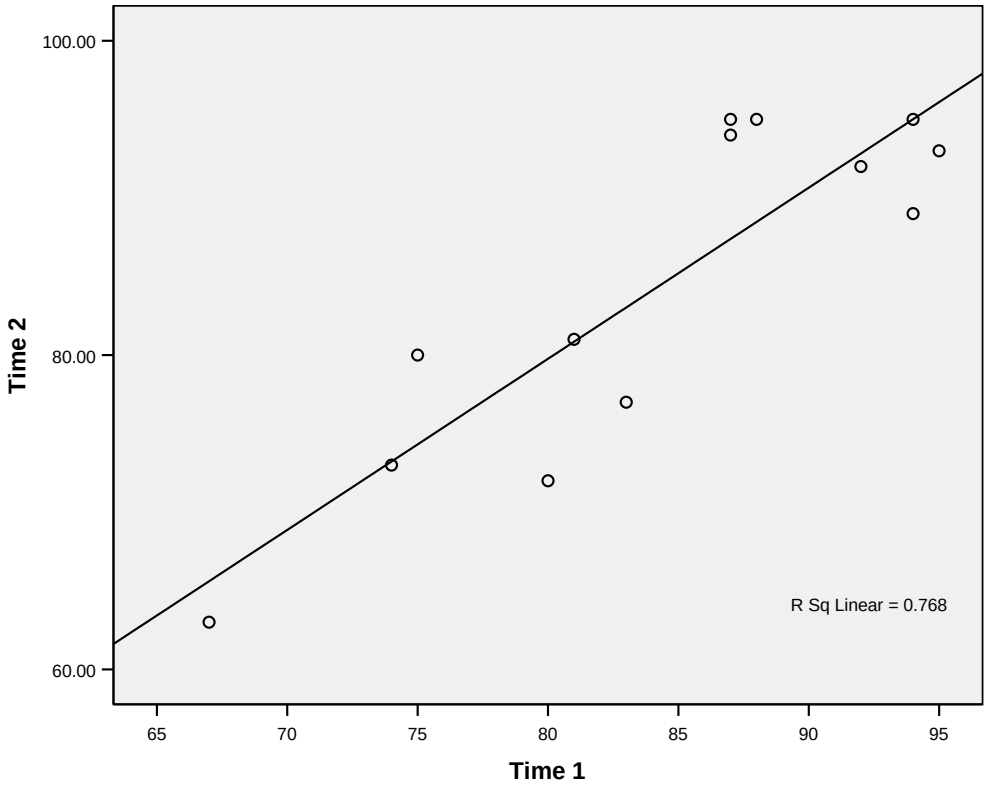
Figure 8-4. Scatter plot demonstrating the correlation between trainer and trainee scores for the DOTS when used in local units



8.3.2.3 Test –retest reliability

The median interval between DOTS evaluations was seven days with a range of one to 35 days. Using all the DOTS evaluations (including both the trainer and trainee evaluations) the test-retest reliability was calculated using Pearson's correlation as r = 0.877 and Spearman's Rho was 0.751; both were significant at the 0.001 level. This correlation is shown by the scatter plot in Figure 8-5 where 76.8% of the variance was stable over time. This gives a SEM of 4.89 with a 95% confidence interval of 9.58.

Figure 8-5 Scatter plot demonstrating the correlation of DOTS scores over time when the DOTS was used in local units



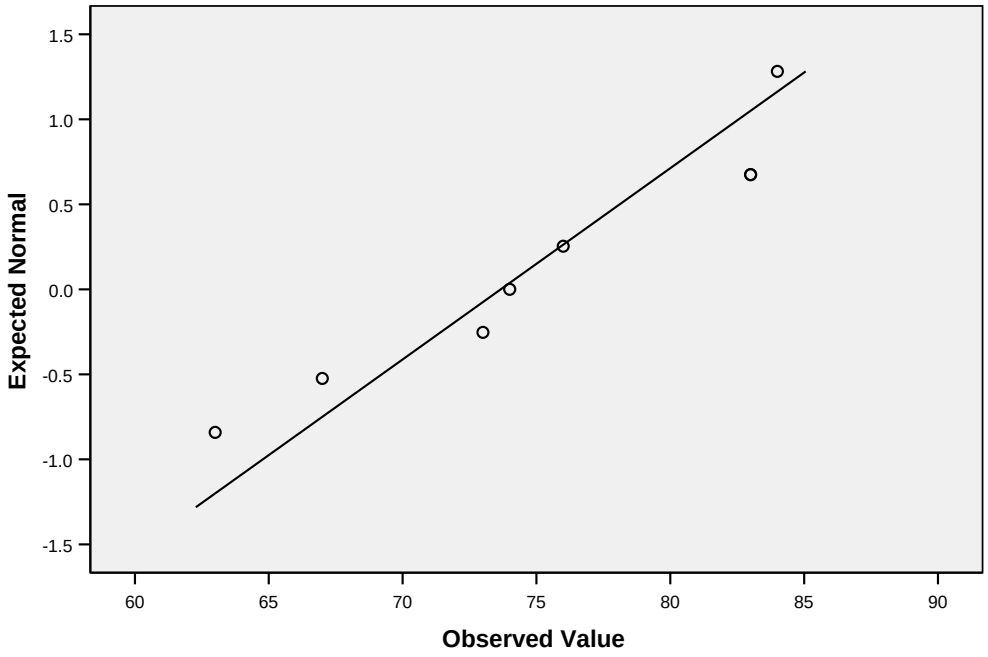
8.3.3 LETS

Descriptive statistics for the LETS data as a whole is seen in Table 8-1. The median score was 74, out of a possible 85 and the range was from 60 to 8 4. The tests for normality are shown in Table 8-4 and the Q-Q plot for normality is shown in Figure 8-6. The Q-Q plot suggests that the data is normally distributed and the z-scores for skewness (z= 0.42) and kurtosis (z=0.92) both were not significant at the $p<0.05$ level. In order to enable comparison with the DOTS again both parametric and non-parametric tests were performed.

Table 8-4. Descriptive statistics for the LETS when used in local units

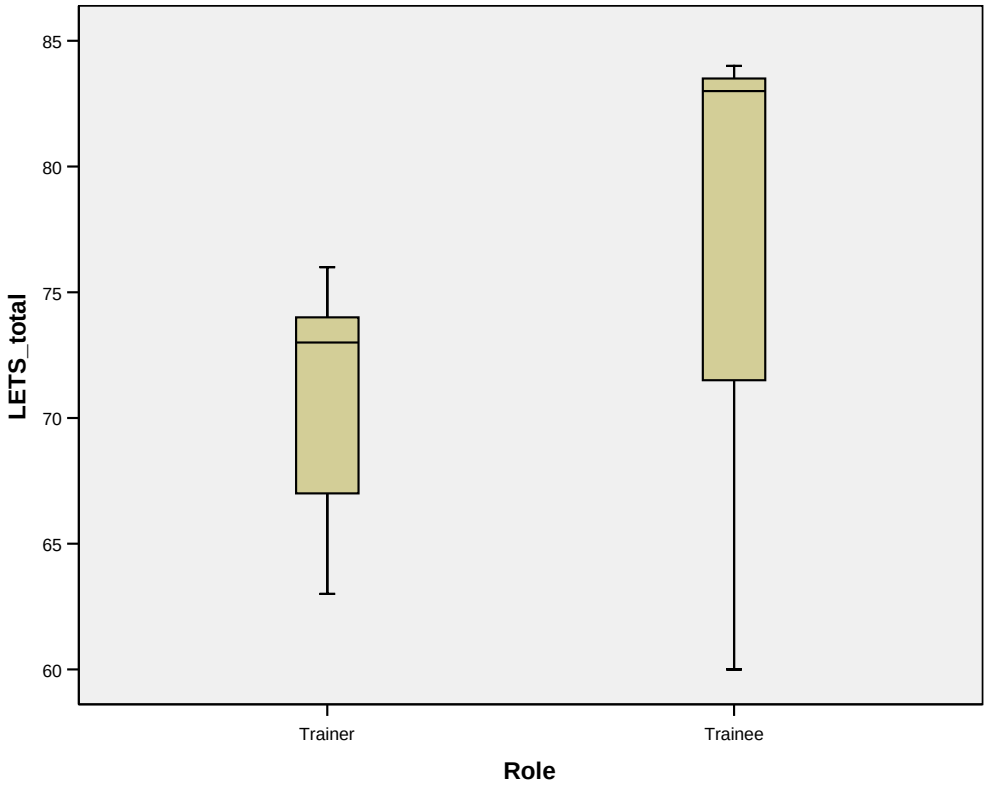
Descriptor	Statistic
Mean (standard error)	73.67 (2.963)
Median	74
Variance	79
Standard deviation	8.888
Minimum	60
Maximum	84
Range	24
Interquartile range	18
Skewness	-0.302
Kurtosis	-1.292

Figure 8-6. Q-Q plot for normality of the LETS total data



Comparing the trainer and trainee data, the boxplot is shown in Figure 8-7, the median score given as a self-evaluation by trainers was 73 with a range of scores from 63 to 76 compared to trainees who gave a median score of 83 and a range from 60 to 84.

Figure 8-7. Box plot comparing trainer and trainee evaluations for the LETS when used in local units



Similarly to the DOTS there was a tendency for respondents to only use the top half of the scale but this was even more pronounced for the LETS than the DOTS, for only three items was the bottom of the scale used by any of the respondents. In considering individual respondents, no respondent gave the same score for every item but clearly there was preponderance by the whole group to use the top half of the scale.

8.3.3.1 Internal structure of the LETS

The overall Cronbach's alpha for the LETS was 0.948, when just considering trainer evaluations Cronbach's alpha was 0.848 and for trainees was 0.978. The corrected item-total correlations and the Cronbach's alpha if item deleted are shown in Table 8-5 for all evaluations. This information is not shown separated into trainee and trainer evaluations because the sample size was very small and therefore it was not possible to examine the internal consistency for the two sub-groups. Considering all evaluations all item-total correlations were greater than 0.3 however if Q7 (taught the whole process of endoscopy) was deleted then Cronbach's alpha was increased.

Table 8-5. internal structure of the LETS when used by trainers and trainees in local units

Items	All evaluations (Cronbach's alpha 0.948)	
	Corrected item-total correlation	Cronbach alpha if item deleted
Q1_made the trainee welcome	.582	.947
Q2_ agreed and worked towards common objectives with a long term plan	.689	.946
Q3_matched their approach and pace to my needs	.739	.945
Q4_ used teaching aids that can support learning	.846	.942
Q5_took advantage of opportune moments to teach	.817	.944
Q6_checked my understanding of the theory of endoscopy	.588	.948
Q7_taught the whole process of endoscopy	.322	.951
Q8_ensured accurate reports were produced	.729	.945
Q9_gave me opportunities to ask questions	.857	.944
Q10_was patient and calm	.775	.944
Q11_was available and focused on me	.832	.942
Q12_developed a good working relationship with me	.857	.944
Q13_set a good example through their own professional behaviour	.797	.944
Q14_built the trainee's confidence	.768	.944
Q15_reviewed the data collected to inform feedback	.696	.946
Q16_helped trainee to reflect on performance	.821	.943
Q17_reviewed long-term progress	.577	.947

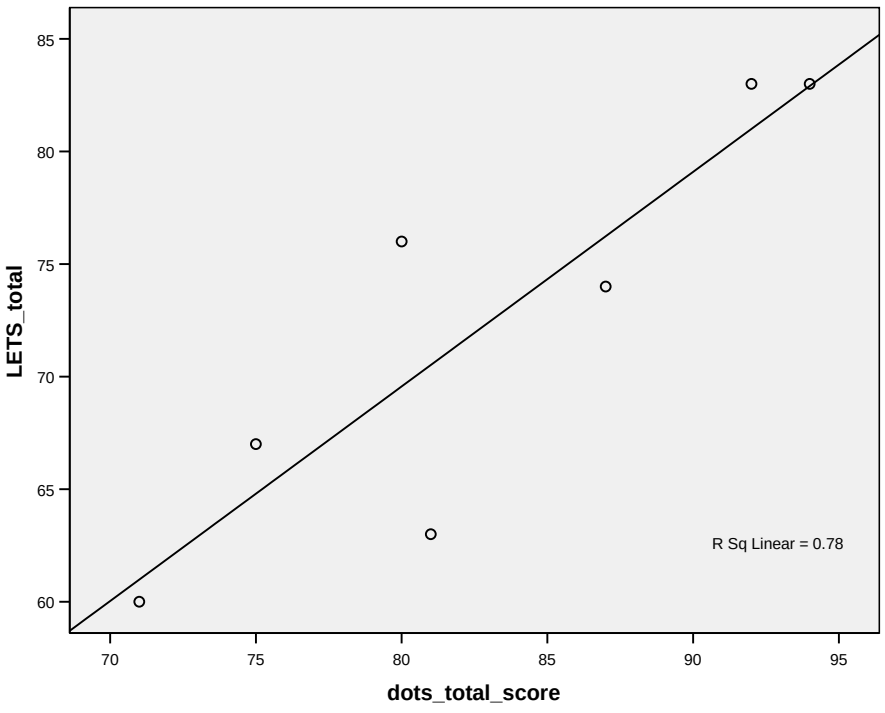
8.3.3.2 Inter-rater reliability of the LETS

In order to consider the inter-rater reliability only paired data could be used. Surprisingly there was a negative correlation with the Pearson correlation -0.636, and the Spearman rho -0.738; neither of these results were statistically significant; this is likely because the analysis was not sufficiently powered

8.3.3.3 Comparing the LETS with the DOTS

Using a two tailed Pearson's correlation the correlation between the LETS and the DOTS was 0.833 ($p<0.01$) and with Spearmans Rho the correlation was 0.811($p<0.05$). This is represented graphically in Figure 8-8. Although this combines both trainee and trainer results this is still a small sample size but the results still reached statistical significance which suggests there was sufficient power for a correlation of this magnitude.

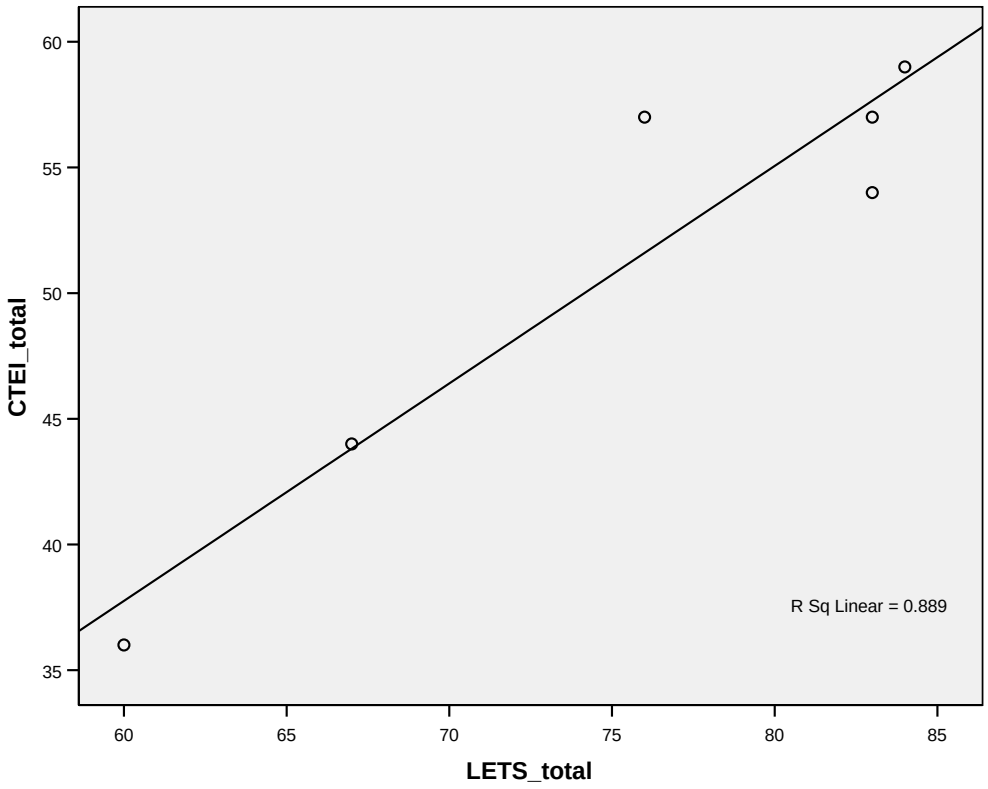
Figure 8-8. Correlation between the DOTS and the LETS



8.3.3.4 Comparing the LETS and the CTEI

Pearson's correlation for the LETS and the CTEI was 0.943 ($p<0.01$) and Spearmans Rho was 0.868 ($p<0.05$); represented in Figure 8-9.

Figure 8-9. Correlation between the LETS and the CTEI



8.3.4 Free text comments

Trainees left more free text comments than trainers. Three trainers made comments on the DOTS and only one trainer made a comment on the LETS compared to trainees of whom ten made comments on the DOTS and four on the LETS. Trainer comments tended to be reflective in nature or gave explanations as to why their training may have been sub-optimal such as time or service pressures. All of the comments made by trainees on both the DOTS and LETS were positive in nature. Often some of the comments made on the DOTS were not necessarily specific to that list but were more general opinions regarding the trainer in question. Many of the trainees commented on several different aspects of the training. Similar comments have been grouped into themes displayed in Table 8-6. All the themes mapped to concepts that were already mentioned on either the DOTS or the LETS.

Table 8-6. Table of free-text comments left on the DOTS and LETS grouped into themes

Allows sufficient time
Busy list but never felt pressurised for time (DOTS)
Allows me to struggle at times which is positive, other trainers can take the scope too early which results in not understanding how to resolve difficulties (DOTS)
Gave me enough time for procedure (DOTS)
Allows me time to do the endoscopy (DOTS)
Allows me time to work out solutions to problems (e.g. looping) before making recommendations (LETS)
Feedback
He has made me aware of how he feels about my scoping i.e. when he's happy with how things are going, which things concern him and how to change if necessary
Excellent training opportunity, gave me enough time for ….feedback (DOTS)
Completes feedback in a timely fashion (LETS)
Constructive feedback, always makes suggestions about future development. Used positive reinforcement effectively (LETS)
Good Constructive criticism (LETS)
Demonstration
Keen to show and demonstrate how loops appear (DOTS)
Explanation
Gave me specific explanations for sequential movements to cannulate CBD by using example of doorframe. Very practical way of explaining things. (DOTS)
Intervention/ Problem solving
I find we both understand now when I require intervention (DOTS)
when I'm struggling asks me why this is and what I could do about it, i.e. encourages me to solve problems before making suggestions (DOTS)
Intervenes appropriately (DOTS)
Excellent at understanding where I have problems (LETS)
Objective setting
Understands what I need from lists at the minute i.e. need to get unassisted numbers up so gives advice without taking over (DOTS)
Clear objectives set and specific skills to work on identified (DOTS)
My trainer always takes the time to understand my training requirements, but is able to adapt this when circumstances change so that all training episodes are utilised
Trainee/ trainer relationship
Very encouraging and appropriate (DOTS)
Very encouraging and supportive (DOTS)
Creates a relaxed environment (DOTS)

Patience
Very patient (DOTS)
Extremely patient (LETS)

8.4 Discussion

Response rate was very low for the base unit study with only a third of those that were sent packs either returning a completed tool or indicating that they currently were not acting as an endoscopy trainer. Asking trainers to complete the tool online was unsuccessful. This may have been because in the endoscopy room there is normally only one computer and this is already being used to create reports for the individual procedures. This requires the endoscopist who performed the procedure to complete the report after the procedure. The computer is also used for the trainee to log their procedures on the JETS website, for the trainer to complete formative DOPS for their trainee and for the trainee to complete the JETS trainer feedback tool. Given all of this the computer may just have been too busy for yet another tool to be completed. The disadvantage of the paper tools was that as it was sent to the trainer's office it required the trainer to remember to take the tools to the endoscopy unit with them. As ethics required for each tool to contain the unique trainer code, in order to supply this to the trainer it was not possible just to leave a collection of the tools within each unit.

Given this very small sample size the results must be interpreted with caution and may not be found to be generalisable to larger groups; however, many of the correlations were statistically significant. An alternative method to engage participants in this study may have been to approach trainees rather than trainers as they are likely to have a greater desire to improve training and therefore may have engaged in the study more fully; this may have improved participant numbers. In future if the tool were to be adopted for general use I would need to consider how to improve engagement. Drivers to participation could include education with regard to the tool, explaining how feedback from the tool would be used to benefit both the trainer and trainee. A driver to participation could also be to make the tool compulsory within units; this obviously however needs to be balanced with ensuring that the tool still remains meaningful as by just making the tool compulsory does not necessarily ensure that those completing it provide useful feedback. Training leads advocating its completion, monitoring this and discussing results of the feedback with individual trainers (discussed further in

section 9.5) may also improve completion. The opportunity for trainees to provide feedback is already a component of Jag's unit accreditation. It could also be used for GMC revalidation. If adopted by JETS then this would form part of the eportfolio system which may also increase its adoption.

Considering both the DOTS and the LETS, although a range of scores was given there seemed to be a tendency for these scores to be at the higher end of the scale. For many of the items, particularly on the LETS only the top half of the scale was used. As only the top half of the scale was used this may suggest an inappropriate choice of scale; as discussed in Chapter 5 the scale could be more positively weighted so that there are more options to the left of neutral; this may increase the range of scores given.

There was also tendency for trainees to rate trainers more highly than trainers rated themselves (Figure 8-3 and Figure 8-7). The fact that learners tend to rate their teachers highly has been seen in other studies(Guyatt, Nishikawa et al. 1993, Beckman, Ghosh et al. 2004) and there is evidence that learners rate teachers higher than peers (Beckman, Lee et al. 2004). The reason for these high ratings may be due to the halo effect, which states that respondents use an overall impression to rate others and either perceive them to be wholly good or wholly bad rather than marking them on specific attributes, leading overall to either very high or very low ratings(Jacobs and Kozlowski 1985, Streiner and Norman 2008). Jacobs and Kozlowski (1985) also found that the halo effect appears to increase with respondent familiarity with the subject that is being rated. This could be argued to have been demonstrated with the LETS where the upper half of the scale was used to a greater extent compared to the DOTS. The other reason that this may have occurred is because some of the items on the LETS are more subjective and more personal than those on the DOTS, due to this the learners may have felt that it was more difficult to disagree with such items. The fact that not all the scale was used makes it more difficult to discriminate between trainers. This is not necessarily an issue as the tool is not summative but it may suggest that trainees are being uncritical and this influences the value of feedback.

The other reason that high ratings may have been seen in this study is because this was a self-selected group of trainers. These are therefore likely to have an active interest in training and may well see themselves as good trainers making them more likely to participate; hence both the high self and trainee evaluations. Given that only

high ratings were used it is not possible to determine from the data whether this was because as trainers were approached only motivated trainers participated who do excel or because the tool's scale is not appropriately scaled because it cannot discriminate between trainers. The other limitation based on these high rating scales is that the results therefore only apply to high ratings and therefore the tools appear to be reliable if ratings are high but does not necessarily mean that the tool would also be reliable for lower ratings, although likely this cannot be proven from the data collected. Approaching trainees rather than trainers initially to participate in the study may have led to a greater spectrum of ability of trainers and therefore a greater spread of scores.

Of note all the trainers in local units were either gastroenterologists or nurse endoscopists (although the trainees completing the tools varied from nurse endoscopists, surgical and gastroenterology trainees). This means than no surgical trainers were evaluated. Although sample size is too small for formal analysis there appeared to be no obvious discrepancy between the nurse endoscopy and gastroenterology trainers; it is obviously not possible to know whether this would have been the case for surgical trainers and it is a limitation of the data that they were not included. It is not possible to know whether the tools would also evaluate surgical trainers reliably, however all types of trainer were assessed by peers and showed no obvious differences between class of trainer therefore this is also likely to be the case when completed by trainees.

In terms of the internal structure of the DOTS, Cronbach's alpha was high at 0.945 when all evaluations were considered. This was perhaps expected as the tool contains a large number of items (as discussed in Chapter 6 and 7). Examining the total data all items showed at least moderate corrected item-total correlations of over 0.3 (Field 2009) indicating that all items measure the same construct. When Cronbach's alpha if an item was deleted was examined, there was evidence that deleting two of the items would increase Cronbach's alpha, these were items 2 (The trainer ensured the trainee knew the role and name of each member of the endoscopy team so that the trainee felt supported) and Item 14 (The trainer allowed the trainee reasonable time to carry out the procedure). Exploring these items further and considering the differences between trainer and trainees, it is noticeable that the item-total correlation for item 2 when evaluated by trainers was low compared to trainees for whom this item correlated highly with the total score. This suggests that trainees are much more

aware of whether trainers do actually introduce the trainee to the team whereas the trainers are much less aware of this occurring. A similar finding occurred for item 14 when trainee and trainer evaluations were considered separately; except the converse was found. For trainers, item 14 correlated highly with total score whereas for trainees this item correlated poorly. This maybe because it is difficult to for trainees to judge what counts as reasonable time and may feel that the trainer takes over the scope too quickly or leaves them to struggle too long. They are very involved in what is going on whereas the trainer is more aware of external factors such as the difficulty of the task or patient discomfort and therefore has a different perspective whether they have left reasonable time. Interestingly item 11 (did not overburden the trainee with too many tasks) although if deleted did not increase Cronbach alpha did show the opposite correlation pattern to item 14 in that it correlated poorly with total score when evaluated by trainers but highly when evaluated by trainees. This may be because a trainee is likely to be very aware when they are being or feel overburdened but it is more difficult for the trainer to judge this.

Test-retest reliability of the DOTS was reasonable with a Spearman Rho correlation of 0.751; this means that over time the DOTS displays reasonable reliability, over 0.7 suggested for formative evaluation (Downing 2004) . One would not expect the reliability to be a hundred percent between the two sessions as these two sessions would not have been the same; Classical test theory does not make it possible to account for other factors within the analysis. For instance each session would have contained different cases with differing levels of difficulty. Other aspects between the two lists such as differing nurses assisting or different time pressures would also have existed but even given these factors the trainer's ability to teach appears to be consistently evaluated by the DOTS over time.

Inter-rater reliability referred to the correlation between trainer and trainee evaluations; this was 0.516, this demonstrates a moderate correlation but not a perfect one. Discrepancy between trainee and self-evaluations has been seen previously in other studies (Claridge, Calland et al. 2003) where 61% of teachers scored themselves significantly differently from their learners. One might hypothesise that this has occurred for several reasons; one argument as discussed above is that learners typically rate teachers highly leading to discrepancy between the two sets of scores. A different argument may be that the teachers are more self-critical of themselves. Schwarz and

Oyserman (2001) suggests that teachers attend to contextual issues and trainees to dispositional; so long term ratings tend to converge as both attend to dispositional issues over the longer term. In terms of endoscopy this means that trainers are more likely to be aware of exactly what happened in that list and complete the tool accordingly whereas a trainee is more likely to respond dependent on what their general opinion is of the trainer. Over the longer term both will respond relying on their overall opinion and therefore scores may correlate more highly which according to this theory we would therefore expect to see on the LETS. Another explanation for this difference in scores may be explained by the differences seen when examining the internal structure of the tool. Clearly there are certain items that seem to be being responded to differently by trainers and trainees, this may relate to how intimately involved they are as a result of a behaviour and therefore have different views on it.

The LETS was not adequately powered to consider differences between trainees and trainers. Provisional results suggest a negative correlation i.e. the higher the trainee scores a trainer the lower the trainer scores themselves. These results are not statistically significant or appropriately powered therefore if the study was repeated with larger numbers this finding may not be replicated. I performed a post hoc power calculation (StatsToDo 2013)based on an expected correlation of 0.636, powered to detect a 0.05 significance at 80% power; in order for this correlation to be sufficiently powered I would require a sample size of 14 pairs. One hypothesis as to why there is a negative correlation may be that the better the trainer is, the more self-critical they are in order to strive to improve further. This would need to be examined using a larger sample size.

With regard to providing evidence of relationship to other variables, there was a high degree of correlation between the DOTS and the LETS and also the LETS and the CTEI. The former of these two correlations suggests that the LETS and the DOTS are measuring a similar construct. This is the expected result as the aim of the DOTS and LETS was that they should both measure the construct of the excellent endoscopy teacher. The fact that the LETS correlates with an already validated instrument is important as it adds evidence that the LETS is measuring the same construct as the CTEI, that of good clinical teaching.

Though the sample size was very small and therefore the results of the analysis must be interpreted with caution the correlations have all still reached statistical significance.

The correlation of the LETS and CTEI is a strong source of validity evidence within the 'relationship to other variables' source (American Educational Research Association, American Psychological Association et al. 1999). Beckman et al. (2004) describe this source as powerful evidence of validity. It must be noted that the CTEI has not previously been validated for endoscopy teaching but it has been previously tested across a wide range of differing specialties and institutions suggesting that its results are not specialty specific and it would likely possess similar validity within this field too.

Fewer trainers completed the CTEI than trainees; this may have been because they had not appreciated that they were expected to. For both the LETS and the DOTS on the trainer self-evaluation the item stem was changed to the first person. I did not do this for the CTEI as I wanted to reproduce it in an identical format in order to preserve its validity; therefore I reproduced it as presented in the paper written by Copeland et al. (2000). Some trainers therefore may have not realised that they were also supposed to complete this tool as well. Including the CTEI completed by trainers may reduce the CTEI's validity as there is no literature surrounding its use as a self-evaluation tool. The analysis was rerun using just trainee evaluations and in fact the correlation between the LETS and the CTEI was even higher (Pearson's r= 0.986 ($p<0.05$), Spearman's Rho ρ =0.949) but with so few cases it did not reach statistical significance when analysed as non-parametric data.

Previous research has shown that teachers find free text comments helpful (Stalmeijer, Dolmans et al 2010). As previously mentioned all the comments made by trainees were in support of their trainer. The themes that arose from the comments, as can be seen in Table 8-6, all concerned concepts that were covered by the items in the DOTS or the LETS. As discussed by Cox et al (2002) this could be because those completing the tools are just re-iterating the items listed above in the tool or that there are in fact a few major characteristics that describe superior teaching. The fact that some of the comments overlapped the tools, for instance patience is included as an item on the LETS but was mentioned by the trainees in the DOTS, suggests that the latter explanation is true. This can be used as further evidence of the content validity of the toolkit as it suggests that as no new themes arose that the toolkit adequately covers the important aspects of endoscopy training (from a trainee's perspective). The two areas that the most comments pertained too were that of feedback and timing, in terms of making sure the trainee was given enough time to try and complete the

procedure independently. The fact that these two areas were noted within the comments section suggests that trainees feel that these are two concepts that are important for a good training experience.

There was some crossover between free-text comments on the DOTS and the LETS. For instance, the interpersonal type items are all contained on the LETS, however several of the comments on the DOTS related to these attributes, such as the fact that the trainer was patient or encouraging. Similarly, the items that relate to how the trainer structures a list in terms of timing are all included in the DOTS but trainees left comments that related to these areas on both tools. This is not totally unexpected as attributes that are shown on a single list are likely to be demonstrated on lots of lists and similarly items that I had considered more global could also be shown in a single list. It highlights that there is a danger in considering the division between the two tools as an absolute. In the Delphi the question asked to the panel was on which tool an item would be most appropriately placed and in round 2 the 'both' option was the most frequently selected item.

One of the trainers commented

> *'I had not thought of introducing the trainee to the staff each week. Will always do this from now'*

This demonstrates how self-reflection can be powerful in helping a trainer think more closely about their training. It also highlights that evaluation tools can also end up being used almost as a curriculum or structure for trainers. This was also picked up by Sarker et al. (2005) as they explained that their eventual aim for their surgical teaching evaluation tool is that their tool be used as part of a program of staff development. This emphasizes that the content of the tool is of key importance as it will influence further training and therefore it is important that it reflects good practice.

All of the free-text comments were of a positive nature. This is likely to be positively affirming for a trainer to read and may positively reinforce attributes they are excelling at in their training. It does not necessarily help improve training further. The tools asked the comments to be as specific as possible and generally the comments were specific in nature, in order to further improve training rewording the comments text may prompt trainees to not just be positive but also constructive. The text in a future version of the tool would be changed to

‘Comments – please be as specific as possible, in particular, consider what the trainer might have done to improve this teaching episode further’

8.5 Conclusion

In conclusion there appears to be good correlation between the DOTS and the LETS and also evidence that the toolkit correlates with other measures of clinical teaching. Results of the DOTS are stable over time with reasonable reliability of the scores; and there is a moderate correlation between trainer and trainee evaluations. There were no discrepancies between the concepts that arose from the free-text comments and those covered by the items in the toolkit although there was some crossover between the two tools. There were some differences in how the items were scored particularly between trainers and trainees; this will be discussed further in chapter 9.

Chapter 9. Discussion

The purpose of this study was to try and produce an evaluation toolkit that could be used to give formative feedback to endoscopy trainers in order to try and improve endoscopy training as a whole. An evaluation toolkit was chosen as evidence exists that giving feedback to teachers can improve their practice (Litzelman, Stratos et al. 1998, Maker, Curtis et al. 2004). The aim was that this toolkit could be completed by trainees, peers and trainers as a self-evaluation exercise. The project has focused on the development of the toolkit and has not evaluated whether giving endoscopy trainers such feedback does make a difference to their teaching practices, which was beyond the scope of this project. The final toolkit consisted of two components the DOTS (Directly observed teaching skills) which could be completed by a peer, trainee, or trainer as a self-evaluation after a single list or procedure and the LETS (long-term evaluation of teaching skills) which could be completed at the end of a rotation by a trainee or trainer. In this chapter I would like to reflect on the process of toolkit development and consider the evidence provided through the toolkit's development of its validity for its use as a tool for the formative evaluation of endoscopy trainers.

During the development of the toolkit I have placed particular emphasis on the American Psychological and Educational Association standards for Educational and Psychological testing (American Educational Research Association, American Psychological Association et al. 1999) particularly in reference to their categorisation of possible sources of evidence for validity. These possible sources of evidence are content, response process, internal structure, relationship to other variables and consequences. This study has focused on the first four of these standards; further work is required to consider the consequences of the use of the tool, the effect of feedback and the tool's impact on both trainers and trainees. As previously mentioned this is not the only way in which validity can be considered but using these possible sources of evidence have helped provide a useful reference guide in considering the process of how the tool should be developed and initially trialled. As well as being a useful guide it has highlighted some further questions about the toolkit. Below I discuss each source and the evidence I have provided but also consider each area's value in providing meaningful feedback to endoscopy trainers.

9.1 Content validity

In order to inform the content of the toolkit I opted to use the attributes described by Wells (2010) who aimed to describe the high-quality endoscopy trainer. Using the list of attributes developed by Wells (ibid) formed a large part of the evidence for content validity. This was because this was a list of attributes that already described the construct that I was attempting to measure. I chose this list of attributes as this concept of high-quality training is one to which I wanted all endoscopy trainers to aspire; therefore these attributes were an appropriate starting point. Also the purpose of Wells' (2010) research was to explore the question of what defines the high quality trainer whereas previous work had concentrated more on the process of learning endoscopy (Teague, Soehendra et al. 2002, Thuraisingam, Madonald et al. 2006) rather than the individual trainer. Wells (2010) had considered the views of different stakeholders whilst developing this list of attributes; the attributes therefore reflected the views of 'expert trainers', those who taught on endoscopy training courses, base hospital trainers, trainees and nurse endoscopists. This includes all those who might use the future toolkit. This also contributed to the content validity of the toolkit as it meant that the toolkit would not just represent one stakeholder group's views on what defines an excellent endoscopy trainer but reflects the views of several groups and therefore creates a greater understanding of the construct in question. The list of attributes was then used in a Delphi process.

9.1.1 Using the Delphi process

The Delphi process was used to allocate the items to the DOTS or LETS. It also reduced the number of attributes as Wells' (2010) list was extensive; there is no ideal tool length but I wanted to develop a toolkit that had utility, i.e. that it was easy to use and not too time consuming. This was later supported by the views of those involved in the Delphi process when several of the panel commented that they felt that it was important that the two tools were short and quick to use.

The items could have just been selected by myself and my research team; however as there was overlap between mine and Wells' supervisory team there could have been influence over what items were included. Using a group technique to gain consensus on the items can be helpful as it is perceived that the pooled intelligence is greater than that of the individuals combined (Clayton 1997). This pooled intelligence can only be

deemed greater though if it is derived from a group of appropriate individuals; the choice of experts is therefore very important in ensuring the consensus has authenticity (Murry and Hammons 1995, Clayton 1997). Experts in this study were taken from four different groups to try and ensure all stakeholders were involved as this is felt to contribute to the validity of the tool. This is important as different groups have previously been shown to have different views on the most important teacher attributes(McLean 2001, Chitsabesan, Corbett et al. 2006). In the surgical tools I reviewed different stakeholder groups were involved in tool development but how their views were incorporated was unclear. The Delphi process has provided a strong transparent method by which this occurred thus strengthening the content validity of the toolkit.

The additional benefit of using the Delphi process is that it involves those who will later be using the tool and that this can create 'buy-in' (Clayton 1997) and create some sense of ownership by the endoscopic community; for this reason all those who teach on the JAG approved training courses were invited to take part. In a previous study adoption of peer evaluation was felt to be more successful when those who would later be involved in peer evaluation were consulted in the development process (Irby 1983) The range of experts I used reflected those used in Wells' (2010) original work which ensures that none of the voices of these separate groups were lost.

Despite concerns that different stakeholder groups might perceive certain attributes to be more or less important the results of the Delphi showed that there was surprisingly little difference in the opinions between the different groups of experts. This suggests that within the endoscopy setting there is little difference between different stakeholder groups views on what attributes describe a high quality trainer. One of the few attributes for which there were differing opinions was the item;

The trainer taught the theory of endoscopy before each new stage

92% of the nurse endoscopists felt that this item should be included in the LETS whereas less than 50% of the panel overall felt that this item should be included in either component of the toolkit. This difference between sub-groups of the Delphi panel may reflect differences in previous training or educational experiences between nurses and doctors who made up the rest of the panel. This demonstrates that it was important to continue to include these different stakeholder groups.

Most of the attributes were positively received by the panel in terms of the general level of agreement. This high level of agreement validates the previous interview work performed by Wells (2010) which further supports the content validity of the toolkit. The level deemed to represent consensus in any Delphi is to some extent arbitrary and different levels of agreement are used in different studies. I opted to use the cut off of 70% agreement as this was similar to other studies (Okoli and Pawlowski 2004, Villiers, Villiers et al. 2005) and I wanted to ensure that the tool contained an adequate amount of detail for its intended formative use. This level was later raised to 77% as a response to missing data. Given the already high level of agreement I felt that this still meant that the tool contained enough detail but clearly some detail was potentially lost; this demonstrates how raising or lowering the level deemed to be consensus can alter the content of the tool and could be considered to be a weakness of the Delphi methodology

Using Wells' (2010) list of attributes has contributed to the content validity by meaning that the items on the toolkit measured the intended construct; the Delphi process continued to contribute to content validity by ensuring that those within the endoscopic community also agreed that the most important attributes were included in the toolkit. One further item relating to ensuring the trainee produced accurate reports was also added during the process. In order for the Delphi process to contribute to the validity of the toolkit participants need to be motivated interested individuals (Clayton 1997). If they are not motivated there is a danger they will not have given adequate consideration to the items and this may decrease the content validity of the final tool. I tried to ensure adequate interest and motivation by asking participants to confirm their interest prior to sending out the first round. Interest appeared to be maintained given the good response rate to both rounds but it is not possible to know how much thought each individual panel member gave to every attribute and some of the high level of agreement may represent the panel just ticking agree for the majority of attributes with very little thought; this is a potential weakness. There is limited evidence that this occurred. Reviewing the Delphi responses only one panel member used only strongly agree and agree but this member did also comment at the end of the Delphi questionnaire that the list was very comprehensive and therefore may have felt that all the items were relevant and should be included. Eighteen percent of the panel only used the top three scores of neutral, agree and strongly agree but all used a mixture of these three response options therefore there is no objective evidence that

any panel member completed the Delphi process giving little consideration to their responses.

9.1.2 Matching the toolkit back to theory

Whilst I felt that the use of a group to select the attributes to be included in the final toolkit furthered the content validity this could also have had a negative effect on the final toolkit. One potential disadvantage was that, although the panel was presented all the attributes, they were asked to rate each attribute individually. In round two all attributes were grouped by theme and it was possible for the panel to see what attributes had already been included from that group but the panel were not specifically asked to refer to this, nor did the groupings have any theoretical stance. This means that the items were largely examined individually and chosen on each item's own merit. Whilst this may add to the content validity as each attribute was assessed by the panel as measuring the right construct the resulting list of attributes may not fit together in any meaningful way.

I discussed in chapter 2 that using a theoretical model can be helpful in ensuring all components of the model are reflected in the final tool (Streiner and Norman 2008). Although Wells described a model as to how the list of attributes fitted together I did not now know whether the items in the toolkit were representative of this model.

9.1.2.1 Wells' model

The central concept to this model of effective endoscopy training is the processes of scaffolding and fading. Scaffolding is the process by which a trainer can enable a trainee to do or achieve more than they would be able to on their own. As the trainee progresses they require less scaffolding and therefore the trainer input slowly 'fades' away. In this way the trainer helps the trainee move from being consciously incompetent to consciously competent as described by Peyton (1998). These are therefore longitudinal concepts that require the trainer to understand each trainee's current level of competence. In the development of the toolkit I did not feel that it would be possible for the tool to capture the dynamic concept of scaffolding and fading within a single list. To capture these concepts the toolkit consisted of both the DOTS, to be completed after a single list, and the LETS, to be completed after a rotation. The longitudinal nature of these concepts has also been noted when medical students considered clinical teaching, they were only able to recognise their supervisors

scaffolding over longer rotations (Stalmeijer, Dolmans et al. 2009). The other advantage to the toolkit containing the LETS relates back to the nature of some of the attributes described by Wells (2010). Some of these attributes whilst important may not be displayed by a trainer on every list either because it is not possible or not necessary to utilise that attribute every time. The LETS meant that such attributes could still be contained on the toolkit. Also as the aim for the DOTS was that it could be completed by peers as well as trainees this may mean that some of the attributes were excluded from the DOTS as it was not possible for a peer to assess however these attributes could still be included in the LETS.

Considering scaffolding and fading, in the LETS one attribute reflected whether the trainer took into account the trainee's current level of competence and the need to match teaching to that trainee's 'learning zone',

> *The trainer matched their approach and pace to the trainee's needs (needs defined by stage, preferred learning style, level of confidence)*

Interestingly, despite the fact that the DOTS was not designed to capture scaffolding, many of the items within the DOTS referred to appropriate scaffolding behaviour. Reviewing the items in the DOTS many of the items referred to the interaction that occurs between the trainer and trainee; this is essentially the skills required within Wells' learning zone. As described in chapter 3, the learning zone is the area that is just outside the trainee's current level of competence and where maximal further learning will occur; it is similar to the zone of proximal development described by Vygotsky (1978). Several of the items relate to the concept of intervention; in scaffolding it may be necessary for the trainer to intervene in order for the patient to have a complete endoscopic procedure but it is important that the intervention is well timed so that the trainee is given the opportunity to perform to their maximum ability i.e. to the edge of the training zone but that the patient remains the priority. The items that reflect this are

> *The trainer allowed the trainee reasonable time to carry out the procedure*

> *The trainer asked the trainee to show where they were struggling*

The first of these items relates to the trainer giving the trainee ample time and the second suggests that the trainee might show where they are struggling rather than the

trainer automatically taking over the scope. Scaffolding is a longitudinal concept but many of the technical teaching skills required to appropriately scaffold a trainee clearly were considered important attributes which should be displayed in every single session.

Whilst one part of Well's (ibid) model considered what the trainer actually does to be an effective teacher the second part considered the attributes required in order to be that effective endoscopy trainer. This model explained how the list of attributes could be grouped into a more meaningful explanation of the good endoscopy trainer. The model is discussed in chapter 3 but briefly to reiterate a six domain model emerged from the analysis of the interviews. These six domains were motivation to teach, ability to develop as a teacher, technical teacher attributes, interpersonal attributes, endoscopy attributes and patient centred attributes. The ability to develop as a teacher refers to the ability to be reflective and be able to evaluate their own teaching in order to improve in the other attribute domains; this toolkit would hopefully complement and aid this process. From the interviews Wells (2010) also felt that in order to be an effective teacher there must be an inherent motivation to teach, often displayed as enthusiasm, which would motivate the teacher to develop their skills within the other attributes but also would directly contribute to effective teaching. Wells' model also highlights that all teaching should occur within a patient centred context as every procedure is performed on a patient for that patient's clinical need, never just for the purpose of teaching.

Wells (ibid) then divided the rest of the attributes into three domains, technical teaching attributes, interpersonal attributes and endoscopy attributes. In order to examine the content validity of the toolkit, I wanted to investigate how much the toolkit reflects these attribute domains.

Table 9-1. Matching the attributes in the Delphi process to Wells' model of the effective endoscopy trainer

Matching attributes from the Delphi process to Wells model
Motivation to teach
Ability to develop as a teacher
Technical teaching attributes
The trainer agreed objectives for the session (either previously or at the beginning of the session) (DOTS)
The trainer ensured the trainee knew the role and name of each member of the endoscopy team before a training encounter so that I was supported (DOTS)
The trainer agreed and applied the ground rules including when to intervene (DOTS)
The trainee questioned the trainee at appropriate times (DOTS)

The trainer provided explanations and descriptions at appropriate times (DOTS)
The trainer used a mixture of suggestions, prompts, solutions and instructions (DOTS)
The trainer checked the trainee has understood instructions and advice by observing or questioning (DOTS)
The trainer used an appropriate quantity of dialogue for me and this teaching episode (DOTS)
The trainer asked the trainee to show where they were struggling (DOTS)
The trainer gave specific skills teaching (examples of this might be keeping luminal view, examine the mucosa, tip control, appropriate insufflation, loop resolution) (DOTS)
The trainer did not overburden the trainee with too many tasks (DOTS)
The trainer demonstrated a procedure where necessary (DOTS)
The trainer intervened in a timely fashion (either at a predefined time or if the trainee was struggling) (DOTS)
The trainer allowed the trainee reasonable time to carry out the procedure (DOTS)
The trainer helped the trainee to assess if the objectives for the session had been achieved (DOTS)
The trainer reinforced positive aspects of the trainee's performance (DOTS)
The trainer identified aspects for the trainee to develop and improve (DOTS)
The trainer agreed and worked towards common objectives during the training period with a long term training plan (LETS)
The trainer used teaching aids that can support learning (e.g. diagrams, the magnetic imager, models etc.) (LETS)
The trainer took advantage of opportune moments to teach during lists (LETS)
The trainer ensured accurate, comprehensive and easily understood reports were produced (LETS)
The trainer reviewed the data collected by the trainee to inform feedback e.g. DOTS forms, CuSum etc (LETS)
The trainer helped the trainee to reflect on the trainer's performance (LETS)
The trainer reviewed the trainee's long term progress (LETS)
The trainer matched their approach and pace to the trainee's needs (needs defined by stage, preferred learning style, level of confidence) (LETS)
Interpersonal attributes
The trainer made the trainee feel welcome (LETS)
The trainer was patient and calm (LETS)
The trainer developed a good working relationship with the trainee (LETS)
The trainer set a good professional example through their own behaviour (LETS)
The trainer built the trainee's self-confidence (LETS)
The trainer was available and focused on the trainee – by minimising distractions (LETS)
Endoscopy attributes
The trainer checked the trainee's understanding of the theory of endoscopy before each new stage (LETS)
The trainer taught the whole process of endoscopy e.g. the indications, consent, communication and sedation (LETS)
Patient Centred
The trainer always ensured that the patient was comfortable and safe and their dignity was maintained (DOTS)
The trainer encouraged the trainee to communicate appropriately with the patient (DOTS)

In Table 9-1 I have allocated the items on the final toolkit under these six domain headings. As can be seen from the table some domains are represented to a greater extent within the toolkit than others; two of the domains are seemingly not represented at all. These two domains were motivation to teach and the ability to develop as a teacher. In terms of the ability to develop as a teacher, as discussed above, the toolkit itself contributes to this domain rather than vice versa. The motivation to teach represents the trainer's enthusiasm to teach but also a desire to improve within the other domains and therefore could be argued to be represented by all the attributes. One could also argue that some of the attributes such as taking opportune moments to teach and making the trainee feel welcome are surrogate markers for motivation and enthusiasm respectively but these attributes could also be grouped within other domains as they have been in Table 9-1.

Two items reflected the patient centred domain and were both included in the DOTS. The fact that both of these items were included in the DOTS accurately reflects the concept that every single teaching episode should be patient centred and therefore should be evaluated every time teaching is evaluated.

Technical teaching contained the greatest number of items and in fact, apart from the patient centred items, all the items on the DOTS were found within this domain. The reason for this finding is likely to be multi-factorial; a similar distribution was found in Wells' original work in that technical teaching also contained the greatest number of items. This meant that in the Delphi there were a greater number of items to select from within this domain. These items also describe discrete behaviours; in the Delphi process several items were combined but this may have been more difficult with these items leading to a greater number being retained. There may also be a feeling among the panel that items within this section are more observable and measurable and therefore were chosen in preference to items that describe less obviously observable behaviours. The fact that the technical teaching items were more highly represented reflects teaching tendencies within faculty development courses in that these also tend to focus on technical skills (Sutkin, Wagner et al. 2008). It may seem that this domain is over-represented and this could be a criticism of the content of the tool. A counter argument is that a key concept of Well's model focused on the learning zone and scaffolding the learner and many of the items describe behaviours that scaffold the

learner. The greater number of these items reflects their relative importance in the model.

Only two of the items on the toolkit could be allocated to the endoscopy attributes domain and in fact when I first allocated these items I felt that they could equally belong to the technical teaching domain. I believe that they describe endoscopy attributes because in order to check the trainees understanding or to teach on the subject the trainer themselves must have adequate knowledge and understanding of the subject of endoscopy. Two of the attributes that were rejected by the Delphi panel from this domain were,

'The trainer demonstrated their competence at endoscopy'

'The trainer has a broad knowledge of the practice of endoscopy'

The first attribute did not meet the consensus criteria for inclusion in the tool. The comments left by the panel regarding this attribute reflected the fact that they agreed that a trainer should be competent at endoscopy but should be evaluated in other ways. The second attribute was initially accepted to the LETS by the panel but was then excluded as it did not meet the generic criteria set by the panel as it was not measurable. Some evaluation tools do ask trainees about the trainer's level of knowledge (Irby and Rakestraw 1981, Litzelman, Stratos et al. 1998, Copeland and Hewson 2000, Conigliaro and Stratton 2010) but a trainee may not be in a position to evaluate their trainer's knowledge. This is a view that is also present within the literature; other studies have opted to not include items that ask learners to pass judgment on their teacher's competence or knowledge as juniors may not be in a position to make such judgments (Guyatt, Nishikawa et al. 1993, Copeland and Hewson 2000). A peer may well be in a position to evaluate another peer's knowledge but I felt that this was not the primary purpose of the tool. Additionally there is evidence that teachers can already find the prospect of peer review threatening(Adshead, White et al. 2006) and I felt that this may be increased if the trainer felt that not only were they going to have their teaching skills assessed but also their endoscopic skills. Trainers are also already assessed on their competence of endoscopy in other ways and are expected to continue to monitor their completion and complication rates (JAG). In summary although endoscopy attributes are part of the model that describes a good

endoscopy trainer they do not necessarily need to be assessed specifically within the toolkit.

The interpersonal attributes were all allocated to the LETS. This may have been because of the way in which they had been worded in that they were all more relevant to an evaluation of teaching over the longer term. A further reason may be that it is more difficult for a peer to judge these attributes therefore they were not allocated to the DOTS for this reason. These interpersonal attributes have been noted to be frequently desirable in a clinical teacher (Sutkin, Wagner et al. 2008), particularly from the view of the learner (McLean 2001, Chitsabesan, Corbett et al. 2006). Many of these attributes refer to what Stalmeijer (Stalmeijer, Dolmans et al. 2009) labels the general learning climate. In a study which considered the learning environment as part of other aspects of learning within the clinical setting(Stalmeijer, Dolmans et al. 2009)students recognised the concept of the learning climate and stated that it was always noted either in a positive or negative way. They referred to it as the way in which they felt welcome, respected and free to ask questions.

I feel that the importance of this interpersonal domain cannot be underestimated. Sutkin et al (2010) in their review of clinical teaching describe the attributes that I would consider as part of this interpersonal domain as non-cognitive attributes. In their review they discovered that over two thirds of the attributes listed within the literature were part of this non-cognitive domain. Other studies also suggest that in fact it is a creation of a positive learning environment that marks out the best teachers rather than just imparting knowledge(Speer and Elnicki 1999).

The interpersonal domain also refers to the relationship between the trainer and the trainee. Many educators feel that this 'teacher-learner relationship is at the heart of the learning experience' (McLean 2001) and therefore is important to capture within an evaluation tool. I feel that it is important to note though that this relationship is not solely due to the trainer alone but refers to the two-way interaction between trainer and trainee. Thuraisingam, MacDonald et al (2006) highlighted the importance of this interaction within their model of endoscopy with both trainer and trainee attributes being a central feature. This is also highlighted by Sutkin et al (2010) who acknowledge that not every learner would desire the same attributes within their trainer and therefore teaching is a more fluid process. Clearly within an evaluation tool of the endoscopy trainer it would never be possible to fully describe this two-way

interaction and the success or demise of such a relationship may not be wholly the responsibility of the trainer. This would be a limitation of any trainer evaluation.

In summary the toolkit appears to reflect Wells' model of the endoscopy trainer although only when both components of the toolkit are used. If a trainer were only to use the DOTS to evaluate their teaching then they would primarily be evaluating attributes within the technical teaching domain.

9.1.2.2 Matching to existing theories

Wells' model (2010) was developed from the list of attributes described in the interviews that he conducted to represent how the attributes interacted and give the attributes more meaning than when examined individually. Using the Delphi has resulted in the patient centred, technical attributes, interpersonal attributes and endoscopy attributes being covered directly but motivation to teach is only included implicitly. Wells' model, although informed by a review of educational theory, was his synthesis to try and describe endoscopy training. His groupings of technical teaching, interpersonal and endoscopy attributes are similar to Sutkin et al's (2010) groupings of clinical teacher characteristics (teacher, human and physician characteristics) and those used by Molodysky et al (2006) to group attributes within the literature but these are groupings rather than an attempt to create a theoretical model. As discussed previously many of the original attributes appeared to map onto the cognitive apprenticeship of teaching (Collins, Brown et al. 1989). As this is an established theoretical model of teaching and the original list of attributes fitted with this model I again wanted to explore whether this was still the case with the final toolkit.

9.1.2.2.1 Cognitive Apprenticeship model

The Cognitive Apprenticeship model (Collins, Brown et al. 1989) is split into four main sections; the content, methods, sequence and sociology of teaching. Rather than just allocate the items to these four main sections I have further split the methods section into component parts. The main reason for this is that this will further distil the attributes within the technical teaching domain and is in keeping with other studies of evaluation tools of clinical teacher that have focused on the methods section of the cognitive apprenticeship model (Stalmeijer, Dolmans et al. 2008)in the belief that this is of greatest relevance to clinical teaching. In Table 9-2 the attributes are listed under the sections and methods described in the Cognitive apprenticeship model; the

methods are described in Chapter 3. The other sections of the model include the content of the teaching, the sequence and the sociology of the teaching. The content and sequence are largely self-explanatory and the sociology of teaching refers to the environment in which learning takes place, the intrinsic motivation of the student and encouraging students to work together (Collins, Brown et al. 1989)

Table 9-2. Attributes in the final toolkit mapped to the domains and methods of the Cognitive Apprenticeship model

Content
The trainer gave specific skills teaching (examples of this might be keeping luminal view, examine the mucosa, tip control, appropriate insufflation, loop resolution) (DOTS)
The trainer ensured accurate, comprehensive and easily understood reports were produced (LETS)
The trainer taught the whole process of endoscopy e.g. the indications, consent, communication and sedation (LETS)
Methods
Modelling
The trainer provided explanations and descriptions at appropriate times (DOTS)
The trainer demonstrated a procedure where necessary (DOTS)
The trainer set a good professional example through their own behaviour (LETS)
The trainer always ensured that the patient was comfortable and safe and their dignity was maintained (DOTS)
Coaching
The trainer used a mixture of suggestions, prompts, solutions and instructions (DOTS)
The trainer reinforced positive aspects of the trainee's performance (DOTS)
The trainer identified aspects for the trainee to develop and improve (DOTS)
The trainer used teaching aids that can support learning (e.g. diagrams the magnetic imager, models etc.) (LETS)
The trainer reviewed the data collected by the trainee to inform feedback e.g. DOTS forms, CuSum etc (LETS)
The trainer reviewed the trainee's long term progress (LETS)
The trainer encouraged the trainee to communicate appropriately with the patient (DOTS)
Scaffolding
The trainer checked the trainee has understood instructions and advice by observing or questioning (DOTS)
The trainer did not overburden the trainee with too many tasks (DOTS)
The trainer intervened in a timely fashion (either at a predefined time or if the trainee was struggling) (DOTS)
The trainer agreed and applied the ground rules including when to intervene (DOTS)
The trainer matched their approach and pace to the trainee's needs (needs defined by stage, preferred learning style, level of confidence) (LETS)
Articulation
The trainee questioned the trainee at appropriate times (DOTS)
The trainer used an appropriate quantity of dialogue for me and this teaching episode (DOTS)
The trainer asked the trainee to show where they were struggling (DOTS)

Reflection
The trainer helped the trainee to assess if the objectives for the session had been achieved (DOTS)
The trainer helped the trainee to reflect on the trainee's performance (LETS)
Exploration
The trainer agreed objectives for the session (either previously or at the beginning of the session) (DOTS
The trainer allowed the trainee reasonable time to carry out the procedure (DOTS)
The trainer agreed and worked towards common objectives during the training period with a long term training plan (LETS)
The trainer took advantage of opportune moments to teach during lists (LETS)
Sequence of teaching
The trainer checked the trainee's understanding of the theory of endoscopy before each new stage (LETS)
Sociology of teaching
The trainer ensured the trainee knew the role and name of each member of the endoscopy team before a training encounter so that I was supported (DOTS)
Not captured by the Cognitive Apprenticeship model
The trainer made the trainee feel welcome (LETS)
The trainer was patient and calm (LETS)
The trainer developed a good working relationship with the trainee (LETS)
The trainer built the trainee's self-confidence (LETS)
The trainer was available and focused on the trainee – by minimising distractions (LETS)

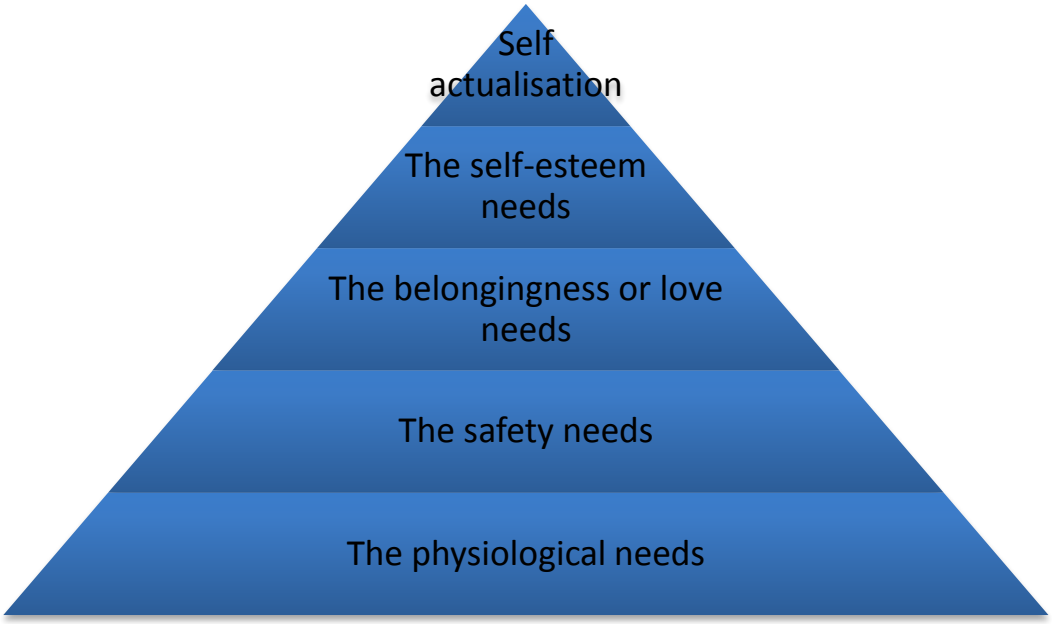
As can be seen from the table all the different domains that are described as part of the Cognitive Apprenticeship model are utilised within endoscopy training and were all represented on either the DOTS or LETS. Allocating some of the attributes was difficult in that it was not always clear under which domain they should be listed. There were also some attributes within the toolkit that did not fit under any of the domains described within the cognitive apprenticeship model. These included many of the attributes that were described as being part of the interpersonal domain within Wells model. This has also been found by other evaluation tool developers (Stalmeijer, Dolmans et al. 2008) who introduced a domain called learning climate to account for this domain. This is a disadvantage of the cognitive apprenticeship model in that it does not take into account the relationship or climate in which the learning takes place; therefore the cognitive apprenticeship model alone cannot describe endoscopy training.

9.1.2.2.2 Maslow's theory of learning

Other theories of learning place a much greater emphasis on the development of the relationship between the trainer and the trainee. Maslow (1970) describes this as a hierarchy of learning needs; his hierarchy is seen in Figure 9-1 (Maslow 1970). His

theory can be described as humanistic (Merriam 2007); it is based on the fact that the learner must have certain needs fulfilled in order for higher learning to occur. The most basic needs must be met first before one can try to fulfil higher needs. The most basic needs are physiological for instance hunger, tiredness; the next are safety needs which include the need to feel safe and secure. Following this are belongingness and love needs which include feeling like one belongs and the giving and receiving of affection. The penultimate needs are esteem needs; learners must have high self-esteem in order to be the right emotional state for learning to take place. Higher learning is referred to as self-actualisation; this will only occur when all other needs are met.

Figure 9-1. Hierarchy of learning needs (Maslow 1970)



If one considers learning of endoscopy using Maslow's theory then the role of the trainer is to try and ensure that all the trainee's needs are met so that learning can occur. Their role is to act as more of a facilitator trying to ensure all needs are met (Merriam 2007). In terms of physiological needs one of the attributes rejected in the first round of the Delphi was

> *The trainer ensured that the trainee was physically comfortable (including neither being tired or in actual physical discomfort)*

The comments regarding this attribute were that it was a little 'precious' and that the trainee must take responsibility for these needs. There were also some comments that

given the current state of the NHS that this may be 'wishful thinking'. One could argue that given the fact that endoscopy trainees are adult learners they should be able to take ownership of these needs. The second set of comments suggest that this may be out of the endoscopy trainers control and therefore not included on the toolkit but may remain a hindrance to training. One of the attributes that was included within the toolkit was

'The trainer is patient and calm'

If a trainer manages to maintain this then their trainee is safe from the anger or impatience of that trainer which would fulfil the trainee's need for safety. Additionally there is an element of safety in knowing that the trainer will scaffold the trainee appropriately and take over a procedure where necessary; this means that the trainee is safe from fear of causing harm to the patient. A feeling of belonging is to some extent measured in the attribute that the trainer makes the trainee feel welcome and develops a working relationship with the trainee. It is also recognised when ensuring the trainee knows the names of others in the team. Although this does not definitely mean the trainee will feel part of the team it may contribute towards a feeling of 'belongingness'. Self-esteem needs are addressed in the attribute

'The trainer builds the trainee's self-confidence'

If one accepts the view that an adult learner should be able to meet their own physiological needs then all the other needs are measured to some extent within the toolkit. Clearly many of the attributes contained within the toolkit do not fit with Maslow's theory of learning; these are often the more technical attributes. The toolkit may not cover all of the concepts at each need level but does seem to reflect some of the underlying principles of Maslow's theory of learning.

9.1.2.3 Wells' provisional toolkit

As mentioned in Chapter 3 Wells' (2010) did create a provisional toolkit for his list of attributes. In order to present the attributes within the toolkit he used the section headings and separate methods described in the Cognitive Apprenticeship model of teaching as listed above. He altered the sociology section to also include elements of preparation for the session. He acknowledged that this model did not cover all the aspects of endoscopy teaching that had been described within his interviews therefore

he also added the categories; learning atmosphere and global. The content domain of the Cognitive Apprenticeship model also contains the concept of teaching heuristic strategies: these are the strategies that are the 'tricks of the trade' which Wells (ibid) used as a separate heading. Wells separated the attributes to the LETS and DOTS using the headings preparation, the learning atmosphere, modelling, coaching, scaffolding, articulation, exploration, reflection and feedback for the DOTS and global, content, heuristic strategies and sequence of teaching within the LETS. The attributes listed under these headings are listed in Appendix 1.

Comparing this provisional toolkit with the toolkit derived from the Delphi process there are some similarities. All of the category headings are represented by items within the current toolkit as would be expected given that I have already determined that all of the Cognitive Apprenticeship domains are captured. There are some interesting differences though between what Wells included in the DOTS and the LETS and the attributes that the Delphi panel included. Wells' DOTS tool contained a 'learning atmosphere' domain; this contained all the items that I have largely discussed above as interpersonal which the Delphi panel placed within the LETS. In contrast in Wells' tool the content domain was included in the LETS whereas the items that have been included from this section were, with one exception, all included in the DOTS. Wells likely contained much of the content within the LETS as he listed each different skill such as keeping the lumen in view, and appropriate insufflation as separate items, and it would not be possible or appropriate to teach every skill on every list. During the Delphi process these separate attributes became condensed into one single item, which referred to the trainer giving specific skills teaching, and used the separate attributes as examples. This condensed item was then moved to the DOTS; this suggests that the panel felt that although not every skill should be taught on every list the trainer should try to deliver some skills teaching every time. This demonstrates that although the domains are adequately represented within the toolkit derived through the Delphi process there are subtle differences with regard to item placement.

9.1.3 Free-text comments

As discussed in Chapter 8 the free text comments left by trainees when the DOTS and LETS were trialled in local units mapped to the concepts that were already covered by the items in the toolkit and no new concepts arose. Most of the comments reflected either the concepts of feedback or providing the trainee with appropriate amount of

time to complete the procedure. Appropriate intervention and the concept of patience were also commented on. The fact that the concepts that arose in the free text comments mirrored those in the items can be used to support validity as no new themes arose. This suggests that the items cover those areas that are important to trainees.

9.1.4 Comparison to the JETS Evaluation tool

In Chapter 2 I discussed that there already existed an evaluation tool for endoscopy trainers that can be found on the JETS website (JAG 2012) and is shown in Figure 2-1. This is a nine-item tool that could be completed after a single list. It was created to fill a need for trainer evaluation but has never been psychometrically tested. One of the options for this study would have been to use the JAG tools and compare all of Wells' attributes against them. Advantages of this is that certainly the tool designed for trainees is in widespread use through the JETS website. The items however were developed using a nominal-group technique by members of the JAG committee who were interested in training (personal communication from JRB). The items were not based on any particular theoretical model. For these reasons I opted not to use this tool to gain evidence for relationship to other variables when trialling the DOTS or LETS within local units. Attempts were made to try and source the data in order to explore the reliability of these tools which could then be compared on a general level however we were unable to obtain this data. It is still worthwhile when considering the content validity to compare the content of both tools. The JAG tool contains items that relate to ensuring appropriate timing or intervention and feedback, which are also contained in the DOTS and appeared important to trainees when considering the free-text comments. The only item within the JAG tool that could be argued to not be covered by either the LETS or the DOTS refers to team working. Although the DOTS makes reference to knowing the name of the other staff it makes no reference to team working. One of the items excluded by the Delphi process was

'The trainer taught the trainee to communicate with the nurses'

which is the most closely related item. Otherwise the content of the two tools is broadly similar, although the items on the DOTS and LETS are more specific and contain greater detail than those on the JAG tool and so gives more precise detailed formative feedback.

9.1.5 Summary

Stakeholders have been actively consulted throughout the toolkit development process Through the use of Well's original interviews and the Delphi process. There was a high level of agreement with all the attributes within the Delphi process and very little difference between the stakeholder groups. Arguably I could have set the level of agreement within the Delphi process at either a higher or lower percentage and this would have meant that either less or more attributes would have been included in the final toolkit. 70% was chosen as it was similar to other studies and seems to have led to adequate coverage of the main theoretical constructs within the tool.

A concern using the Delphi process to select the items was that as the items were all reviewed individually that they might not form a cohesive list, nor might they accurately reflect any model or theory. Others have observed that for evaluation to be useful it should have strong theoretical underpinnings; if there is no theoretical underpinning then it is less likely to bring about improvement in teaching (Bowden and Marton 2000). In considering Wells' model of effective endoscopy training the key concept of scaffolding is reflected in both tools. This was by a more generic item within the LETS which concerns matching teaching to training needs whereas in the DOTS scaffolding is reflected by several attributes that measure behaviours required to effectively scaffold a trainee. The domains that explain how the attributes interact are also reflected in the final toolkit but only when the whole toolkit is used. The patient centred attributes are contained within the DOTS and interpersonal attributes are only contained within the LETS. When I compare the items with Wells provisional toolkit the domains he used are all represented but there are differences between what he felt should be in the DOTS and the LETS and what the panel felt.

In terms of established theories of learning and teaching the toolkit does also reflect aspects of these theories. The methods suggested by the Cognitive Apprenticeship model of teaching are all captured within the toolkit as were all the broad domains. The Cognitive Apprenticeship model does not account for many of the interpersonal skills. I feel that it therefore considers many of the skills used within the learning zone of Well's model but it does not consider the 'glue' of the interpersonal attributes that adhere the trainee to the learning zone. Maslow's theory of learning, which focuses on the fact that the learner must be ready to learn, is largely captured within the toolkit

but in contrast to the Cognitive Apprenticeship model it does not account for the technical teaching attributes. Neither of these theories of teaching or learning fully explain the teaching of endoscopy therefore although they are reflected in the toolkit alone they do not explain all the attributes listed. Similarly although I have discussed how attributes can be matched to domains within these theories this does not mean that each of these domains are fully captured in the toolkit rather one part of that domain is covered.

I have mentioned that to fully represent Wells' model both the DOTS and the LETS tools must be used. This is because the interpersonal domain is only captured by the LETS, which would be completed after a prolonged period of contact with a trainer. In many endoscopy units a trainee is attached to a trainer for a prolonged period of time this enables the relationship between trainee and trainer to both develop and be measured by the LETS. In some units however trainees are trained by a variety of trainers and they would not be able to complete a LETS for a specific trainer. Whilst this might increase the accessibility of training lists the trainee would not necessarily develop a relationship with a trainer and this relationship would not be measured but does this matter? Clearly these interpersonal skills have been incorporated by Wells (2010) and therefore arose from the views of what makes a good trainer of those that he interviewed. Thuraisingam et al (2006) also emphasised the importance of interpersonal attributes of the trainer in order to create a successful learning experience for the endoscopy trainee. Above I have discussed the importance that others have put on the interpersonal skills of a clinical teacher (McLean 2001, Molodysky, Sekelja et al. 2006, Sutkin, Wagner et al. 2008).

Kilminster and Jolly (2000) refer to these interpersonal attributes as contributing to the relationship between trainer and trainee within clinical supervision. They state that 'the supervision relationship is perhaps the most important factor for effectiveness of supervision (p827)'. This relationship would not currently be measured by the DOTS alone. The degree of concern about this arises from whether one feels that an endoscopy trainer's role is to train or supervise the student. When comparing the definitions of supervise and train in the Oxford English dictionary there is little difference with 'to supervise' is defined as 'observe and direct the execution of (a task or activity' whereas 'to train' is defined as 'teach a particular skill or type of behaviour through sustained practice and instruction'. Supervision is defined by Kilminster and

Jolly (2000) as 'the provision of monitoring, guidance and feedback on matters of personal, professional and educational development in the context of the doctor's care of patients (p828)'. Many of those training in endoscopy are likely to have an educational supervisor as part of their parent speciality and therefore one could argue that the endoscopy trainer's role is just to teach the skills of endoscopy however this is not reflected in either Thuriasingam et al's (2006) or Wells'(2010) work on how endoscopy is learnt or taught. The level of supervision would not be acknowledged on the DOTS tool and therefore the tools need to be used in conjunction if one believes the act of supervision is occurring. If, however, it is only necessary to capture the skills training where no supervision occurs then the DOTS alone may be sufficient.

Donnelly and Woolliscroft(1989) found in their study concerning different grades of teachers teaching clinical medical students that some types of teachers did better on the interpersonal function compared to others who did better on the cognitive function. The difference in scores between these different teachers related to the clinical grade of the teacher. The more senior teachers tended to receive higher evaluations within the cognitive domain whereas more junior teachers, who were the residents on the wards, were evaluated more highly within the interpersonal domain. Donnelly and Woolliscroft hypothesise that these differences actually reflect the fact that these different groups of teachers actually have different roles in regard to teaching the students which is reflected in these evaluation differences. The residents have daily contact with the students, giving regular feedback and regular informal teaching whereas the attendings and preceptors have more scheduled teaching with the students once or twice a week which has more specific aims and objectives and therefore the teaching is more formal. This demonstrates that there are different roles of different teachers and that those different teachers will demonstrate different attributes that are accurately identified by learners.

An alternative method for selecting the items would have been to use a nominal group technique. As previously described this requires the participants to meet and individually rank the items. There is then group discussion about the scores given and and through the process of iteration consensus is sought for the final toolkit. This discussion may have led to more debate and the items may have been scrutinised more with the final list reviewed more as a complete toolkit rather than at the individual item level as the items would be discussed in relation to each other. This may have led

to a more cohesive list of items; however I think the concern persists that this may not be the view of the whole group but one or two dominant individuals. I think this would have been of particular concern given the fact that the experts I included in the Delphi came from several different groups and it is possible that some groups may have been dominated by others. In conclusion despite the fact that the items were reviewed and chosen as individual entities the toolkit appears to have retained some theoretical coherence; the fact that these theories can still be applied to the toolkit adds strength to the evidence for the content validity of the toolkit.

9.2 Response process

The response process refers to the methods by which the evaluation is carried out, to try and ensure that the process of completing the tool does not interfere with the respondent addressing the construct in question. One component of this is ensuring that the items are clear and interpretable by all. The two rounds of cognitive interviews have substantially contributed to this source of evidence. Both rounds of cognitive interviewing aimed to ensure attribute clarity and ensure that each attribute's interpretation was unambiguous. This did highlight several issues; for instance, examples were found to be both helpful in improving item clarity but at times limited the respondents to only consider those examples mentioned.

One of the recurring issues within the cognitive interviewing was that of author intent. Several times during cognitive interviewing the respondents would interpret an item slightly differently; this would highlight a potential issue with that item but it would not inform me as to which interpretation was correct. At these times I considered author intent, how did the author of the items intend for that attribute to be interpreted. This was possible as I had access to the author's thesis (Wells 2010) which described how the items had been derived and then written more concisely. I also had access to Wells' original interview transcripts within N-VIVO in which quotations were stored under nodes, which pertain to the individual attributes. This meant that if it were unclear from the longer description of the attribute I could look at the quotations from the original interviews to try and determine the original meaning. This was useful as it meant that I superimposed as little of my own interpretation on the items as possible; but did rely heavily on a presumption that Wells' had interpreted his interviewees' meaning correctly. As I also had access to the interview excerpts contained with each node I was therefore able to also confirm these interpretations.

Use of a Likert scale may also affect the response process. A five point Likert scale was chosen after a review of the relevant literature. There appeared to be no difficulties with the scale when the second round of cognitive interviewing occurred, although this was a very small sample size. Scale choice as discussed in Chapter 2 is often poorly reported in the development of evaluation tools; this is likely because implementing a tool is resource intensive and therefore there is not always the opportunity to trial different response options(Schwarz and Oyserman 2001). One evaluation tool that's scale changed during the pilot process (Conigliaro and Stratton 2010) moved from a dichotomous scale to a three point option as this provided more detail. I hypothesised that a five point scale would provide enough detail without being too labour intensive; however it is difficult to empirically prove that this was the correct scale choice within the timescale of this study. Within the second round of cognitive interviewing all the response options were utilized by each interviewee which suggested that the scale was appropriate and would be used in the same way within the trial setting.

When I examined how the tools had been used during the trial period, different groups used the response processes differently. Considering the DOTS, peers tended to use more of a spread of scores along the scale compared to trainees and trainers. In particular trainees tended to favour the top half of the scale, for seventeen of the nineteen items trainees only used the top three points on the scale; neutral, agree and strongly agree. A similar finding was found with the LETS in that for both trainees and trainers the bottom half of the scale was used for only three items. In Chapter 5 I discussed that increasing the number of points on a scale increases the reliability of the scale(Streiner and Norman 2008) and this was one of my justifications for using a five point scale. The reality though is that for many of the items trainers and trainees only used a three point scale as they did not use the bottom half of the scale; thereby there is a danger that the reliability is reduced. An option to overcome this would be to move the anchor points so that neutral is not in the middle of the scale but is moved towards the left. For instance if I were to do this and still want to use a five point scale the items could read disagree, neutral, and then there would be three positive anchors to the right. These three anchors could read somewhat agree, agree and strongly agree but this would rely on the fact that a respondent could discriminate between the three positive anchors.

It is also worthwhile noting that peers did use the lower end of the scale. There are several reasons why the peers may have used the bottom half of the scale. In other studies peers have been noted to give lower mean scores compared to learners (Beckman, Lee et al. 2004) which may have been reflected in this study with peers using the lower half of the scale. The peer data was also collected on a different trainer group through the 'Training the trainer' courses. Although all the trainers on these courses have opted to attend this is often at the encouragement of the training leads within their units. Although they could opt not to be evaluated none did so. This is in contrast with the trainer and trainee data, which was very much, self-selected and given the poor response rate I could hypothesise that only motivated trainers took part. This might therefore be why only the top half of the scale was used, in that perhaps these actually were just very good trainers. I therefore feel that trialling the tool has raised some questions about whether the scale is appropriate but there is not yet enough evidence to state that it definitely should be altered but requires investigation with further data collection in a non-self-selected group.

One of the other considerations within the response process is the nature of the items themselves. Above I have discussed that I wanted them to be interpreted the same by all respondents and this was investigated using cognitive interviewing but when considering how they mapped back onto theory has caused me to look more closely at the items. My own criteria and the one I requested that the Delphi panel use was that the item be measurable, however I did not stipulate whether this had to be a measurable observation or a measurable judgment. When asking respondents to make judgments these can be referred to as high inference attributes. Inference refers to the process between what is seen and heard and the cognitive or social judgment that is placed upon it (Rosenshine 1975). High inference behaviours are therefore those that are more subjective such as enthusiasm or rapport whereas low inference behaviours tend to be more observable. Many of the attributes within Wells interpersonal domain can be referred to as high inference and therefore may have been less likely to have been agreed with by the panel where one of the inclusion criteria given to them was that the attribute be measurable. This may not have been the case as some of the attributes that have been included remain high inference but may be why this domain is less well-represented on the toolkit. An example of a high inference attribute included on the toolkit is

The trainer developed a good working relationship with the trainee

A criticism of such an attribute may be that it is subjective (Conigliaro and Stratton 2010)and also provides the trainer with little information about how to change were they to be scored poorly on such an item . This is not to say that such an attribute may not be measured reliably, Sutkin et al (2010) use enthusiasm as an example, they argue that although we may all recognise such an attribute within good teachers we have experienced we cannot say how it is done. Conigliaro and Stratton (2010) tried to overcome this by creating a toolkit that only contained observable behaviours however they conclude that this might ignore some of these intangible behaviours that represent this interpersonal domain and ends up focusing more on technical teaching. In trying to capture both the technical and interpersonal domains the toolkit has ended up with a mixture of both high and low inference statements. Whilst some might criticise the degree of subjectivity of some of the items I feel on reflection that their inclusion was important in order to adequately reflect the interpersonal domain.

As well as the completion of the toolkit, response process also refers to how data gathered by a tool is given as feedback to the subjects. In this study I have concentrated on the development of the toolkit, an assessment of whether trainers appreciate such feedback and whether it makes a difference to their training was beyond the scope of this study. It is worthwhile however considering what the product of the toolkit would be and how it should be presented to trainers; although the method by which this data is presented forms part of response process I have chosen to discuss this more fully within the consequences category later in this discussion.

9.3 Internal Structure

9.3.1 Internal consistency

Internal structure refers to how the items relate with each other in terms of the structure of the tool and the toolkit's reliability (Downing 2003). I shall first consider the internal consistency of the two tools. As discussed in Chapter 6 the internal consistency of a tool represents whether all the items on the tool are measuring the same construct; if they are then there should be shared variance between the items and items should correlate with each other. This correlation is presumed to be the degree to which each item measures the construct under investigation (Beckman, Ghosh et al. 2004). In all of the situations that the tools were used they demonstrated

good internal consistency as measured by Cronbach's alpha. The DOTS tool demonstrated a Cronbach's alpha of 0.895, 0.907 and 0.934 when used by peers, trainers and trainees respectively. This suggests that the DOTS tool has high internal consistency and all the items are measuring a similar construct further supporting the Delphi selection process. This high Cronbach's alpha must be considered with caution as alpha is also a function of the number of items that the tool contains (Clark and Watson 1995). This means that as the number of items in a tool increases then Cronbach's alpha also increases without the items necessarily correlating any better. As the DOTS tool contained 19 items it is not possible to just state that as Cronbach's alpha is high that the DOTS tool demonstrates good internal consistency. One way therefore to examine the internal consistency further is to examine each item separately and consider what happens to Cronbach's alpha if that item were deleted (Field 2009). If when an item is removed from the calculation Cronbach's alpha increases this suggests that this item does not share as much variance as the other items and therefore may be measuring a different construct. Each item can also be examined using item- corrected total correlation which examines how well each item correlates with the total score. Acceptable levels of item-corrected total correlations vary from 0.2 (Streiner and Norman 2008) to 0.3 (Field 2009). When this was examined using the peer data all item-corrected total correlations were acceptable (range 0.416-0.652) and Cronbach's alpha was always lower if an item was deleted which suggests that all items contributed to the internal consistency of the tool.

When examining the internal consistency for the DOTS when it was utilised by trainees and trainers it was possible to compare the data between the two groups directly as it was collected on the same set of trainers on the same occasions. In order to examine the internal consistency I combined the data and examined the internal consistency of the tool using all the evaluations and then considered the two groups separately. This gave a Cronbach's alpha for the whole tool of 0.945. When examining the items separately all item –corrected total correlations were acceptable but, in contrast to when it was used by peers, there were items that when deleted the Cronbach's alpha was higher. The items that were a concern were

> *The trainer ensured the trainee knew the role and name of each member of the endoscopy team before a training encounter so that I was supported (Q2)*

The trainer allowed the trainee reasonable time to carry out the procedure (Q14)

When these same two items were examined using trainer and trainee data separately this highlighted differences between the two groups. Item 2 correlated poorly with the corrected total for trainers (0.09) and the Cronbach's alpha if item deleted increased substantially for the trainer data. In contrast when considering the trainee data item 2 correlated highly with the corrected total score and Cronbach's alpha was smaller if this item was deleted. The opposite was true for item 14, this correlated less well with the total for trainees than trainers and the Cronbach's alpha increased when trainee data alone was considered. When considering the data as a whole if any of the other items were deleted Cronbach's alpha decreased, however when trainer and trainee data was examined separately there were more discrepancies within the trainer and trainee data. If item 11 was deleted Cronbach's alpha increased and item-corrected total correlations were low when only the trainer data was examined

The trainer did not overburden the trainee with too many tasks (Q11)

For the trainee data Cronbach's alpha was raised if item 8 was deleted but item-corrected total correlations were reasonable

The trainer used an appropriate quantity of dialogue for me and this teaching episode (Q8)

These differences suggest that trainers and trainees use the toolkit differently. Item 2 refers to ensuring that the trainee knows everyone within the team but also that they are supported; this may correlate less well with the trainers' scores for several reasons. It may be because they are uncertain whether their trainee does feel supported and therefore their responses are more variable. Alternatively they may perceive this measures a different construct of whether the trainee is supported by the rest of the team; trainers may therefore answer in respect to that rather than their own efforts to create a supportive culture within the endoscopy room.

Considering the other items the differences between item 11 and 14 are interesting. Overburdening did not correlate well with the total score when considered by trainers whereas item 14 which refers to being giving enough time correlated poorly in trainee evaluations. I feel that both of these items refer to appropriate scaffolding of the

trainee and giving trainees the correct amount of time and subject matter for them to effectively learn. The fact that these items differ in their correlation between trainees and trainers I feel represents the differing positions of teaching and learning and how these items have been phrased. Feeling overburdened to me suggests a subjective attribute and is personal to the trainee as to whether they feel overburdened or not; making it very difficult for trainers to judge. It is likely that no trainer would intend to overburden their trainee therefore the scores for this would be more variable compared to the other answers. In contrast it may be difficult for the trainee to be able to judge what 'reasonable time' is as they are too involved in the moment of the procedure. They therefore may feel that their trainer takes over the scope too quickly or that the trainer leaves them to struggle for too long but the trainer is perhaps able to make a more objective judgment. They may use the trainee struggling as a method of exploration and enabling the trainee to try different approaches to see what works. Conversely they may be aware that the procedure being attempted is too difficult for the trainee at that stage or the patient is in discomfort and therefore take over the scope earlier. This might explain why these items correlate differently for trainers and trainees.

If an item correlates poorly with the total or if Cronbach's alpha becomes higher when that item is deleted then it is convention that that item should be considered for deletion in order to increase the internal consistency of the tool (Field 2009). Given that these items appear to reduce the internal consistency one could argue that they should be removed from the toolkit. When the Delphi panel were asked to consider the items they were given certain criteria, these were that the item were measurable by all and represent high-quality endoscopy teaching. Although these items were felt to be potentially measurable by all, these different correlations suggest that some may be in a better position to measure than others or be measuring from a different viewpoint. I feel that given the fact they correlate well with one type of evaluator they should not necessarily be deleted as they provide valuable information from that evaluator group. The contrasts between these different groups may provide useful information back to trainers, as previously discussed, discrepancies between student and trainer ratings can be a motivator for reflection and change (Cohen 1980, Stalmeijer, Dolmans et al. 2010). Schuwirth and van der Vleuten (2006) also caution against making decisions based on psychometric statistics alone; they caution that by removing items based on poor correlations alone can lead to the loss of information.

This can be seen here, although there are differences in how trainees and trainers respond to the above items that does not necessarily mean they should be deleted but these differences considered and explored further. This data is also from a small sample size therefore it would be worthwhile collecting more data to see if these findings were replicated in a larger sample group.

It does however show a potential disadvantage of using the same items for all three groups; I opted to do this because I felt that using the same tool allowed direct comparison between items and therefore would more clearly highlight discrepancies between scores but some of the items may in fact have been worded in such a way that they are better answered by different groups. In Section 9.2 I discussed the concept of measurability of the items in terms of them requiring respondents to make high or low inference judgements; these internal consistency results suggest that the degree of inference is also affected by one's role within the teaching episode.

Using real life endoscopy unit data enabled the tool to be completed by both trainers and trainees and enable the above comparisons to be made. Given the above it is interesting to note (although not collected on the same sample population) that for the peer data all items had acceptable item-total correlations and all items contributed to the high Cronbach's alpha, this suggests that the peer, as he is not directly involved within the teaching episode, appears to make judgements across the items with greater consistency.

When the peers completed the DOTS items 12 and 13 were excluded from the data analysis as trainers were told not to take over the scope but it is worthwhile noting that when the tool was used by trainees and trainers these items did contribute to the internal consistency of the toolkit and showed good item-corrected total correlations.

Due to the fact that the sample size was small and there was preponderance to only use the top half of the scale it was not possible to examine each item individually for the LETS but the tool overall did appear to show reasonable internal consistency (Cronbach's alpha 0.948).

Examining the internal consistency gives information on the homogeneity of the tool (Streiner and Norman 2008); however, although all the items appear to measure the same construct, within that construct there may be different dimensions. For instance, earlier, when considering the groupings of items as part of the content validity, the

items were grouped differently in the different theories of what makes an effective teacher even though they all still add up to describe the effective teacher. It is also possible to see if there are sub-dimensions within the scale statistically; this can be performed using factor analysis (Field 2009). It was only possible to perform factor analysis when the DOTS was trialled by peers. This was because factor analysis requires a minimum sample size which was not met when the DOTS or LETS was used by trainers and trainees. In terms of carrying out the factor analysis there are several different methods by which the number of factors can be calculated(Costello and Osbourne 2005, Field 2009). Examination of the scree plot suggested that there was only one factor however when selecting all factors with an eigenvalue of over one then a four factor model emerged. These four factors are shown in Table 9-3.

Table 9-3. Factor structure when the DOTS was used by peers

Factor 1
Q4_questioned the trainee
Q5_provided explanations and descriptions
Q6_used a mixture suggestions, prompts, solutions and instructions
Q7_checked the trainee had understood instructions and advice
Q8_used an appropriate quantity of dialogue
Q9_asked the trainee to show where he or she was struggling
Q10_gave specific skills teaching
Factor 2
Q15_ always ensured the patient was comfortable and safe
Q16_encouraged the trainee to communicate appropriately with the patient
Factor 3
Q11_ did not overburden the trainee with too many tasks
Q14_allowed the trainee reasonable time to carry out a procedure
Factor 4
Q1_agreed objectives for the session
Q2_ensured the trainee knew the role and name of each member of the endoscopy team
Q3_agreed and applied the ground rules
Q17_helped the trainee to assess if objectives for the session had been achieved
Q18_reinforced positive aspects of the trainee's performance
Q19_identified aspects for the trainee to develop and improve

These factors represent groupings of items that appear to have shared variance and therefore statistically appear to be measuring a similar construct. In Chapter 7 I discussed the statistical strengths and weaknesses of this model but in reviewing the items there does seem to be some coherence between the items. Interestingly the attributes in the final toolkit that I felt represented Wells' (2010) patient centred domain are also represented in the factor structure as a separate factor (factor 2). The

rest of the items that made up the technical teaching domain are represented by three different factors. One of these factors (factor 3) contains items that are directed at appropriate scaffolding of the trainee, another factor (factor 1) contains the verbal interactions that occur within the learning zone that help support the trainee. The final factor could either be argued to contain those items that structure the session and relate to opening or closing the session. An alternative argument is that these items represent ensuring the training is trainee centred in making sure the trainee is aware of the point of the session and then reflects back on the trainee at the end.

There are statistical weaknesses within this factor structure; two of the factors only contained two items each which makes them potentially unstable (Costello and Osbourne 2005) meaning that if the data was collected again then these factors may not emerge. If I were keen to use this factor structure, in order to make these factors more stable I would need to add more items that would likely fit into these smaller factors in order to strengthen them The other potential weakness is that the factors that emerged tended to group items together as they appeared on the toolkit, with items next to each other within the same factor. The resulting factors may just therefore represent order effect rather than the explanations I have put to them. This could potentially be examined by retrialling the tool but randomising the items each time and investigating whether the same factor structure emerged.

To determine whether this factor structure was stable and these groupings were consistently seen in the data I could either collect more data and then perform the same exploratory factor analyses as used in this study and see if the same factors emerge. Alternatively I could collect more data and use confirmatory factor analysis (Streiner and Norman 2008), this is when the factors are stated prior to running the analysis and then the analysis is run to see if the data does fit the proposed factor structure. In this case the factors that I have described above would be used as the factors suggested pre-analysis.

9.3.2 Reliability

As well as considering the structure of the tool the internal structure also considers the evidence that the tool is reliable. As discussed in Chapter two reliability is important for validity because if the tool produces entirely different results every time it is used or when it is used by different people then the results become meaningless. In order to consider the reliability of the toolkit I had to balance testing the tool in such a way to

examine reliability as fully as possible whilst trying to trial the tool in an ecologically valid method as possible. In order to overcome this I opted to trial the DOTS and LETS with trainees and trainers as self-evaluations within local units and to trial the DOTS completed by peers on Training the Trainer courses. This has given useful insights into how the tool is used and its reliability under these test settings but it does make it difficult to draw conclusions between peer evaluation and trainee and trainer evaluation.

Considering when the tool was trialled in local units first. This setting had the greatest ecological validity in that I tested the tool in the environment that I wished for it to be used. Collecting enough data within this setting was challenging in that it was difficult to engage those within local units to use the toolkit. This was a frustrating process but is informative when considering how the tool might actually be used in future practice and an area of further work may be to investigate what barriers there were to its use. A requirement of ethical approval was that trainers were used as point of contact in order to ensure that they had consented to be evaluated, however as part of this study they received no feedback from how they had been evaluated therefore there was no personal gain to completing the toolkit or asking the trainees to do so. There is also already an evaluation tool that can be completed by trainees to give feedback to trainers, it may be that having the two tools was too much to be completed at the end of the list or that trainers may feel that the tool already in use is fine and therefore there was no need to introduce or trial a new tool. Lack of response may also indicate a lack of interest or belief that evaluation tools make any difference and as previously discussed in Chapter 1 there is conflicting evidence with regards to their effectiveness in changing teaching behaviour. If trainers and trainees feel that an evaluation tool will make no difference then there will be less interest in trialling the tool. Centra 1993 suggests that in order for significant improvements in teaching to be made as a result of evaluation then four conditions must be filled. These are that the evaluation must provide new knowledge, trainers must learn something new about their teaching; trainers must place value on the evaluation; there must be motivation to change on behalf of the trainer and there must be some understanding how to change as a result of the evaluation. These conditions may also relate to the uptake of evaluation as well as any change that results from it. The poor uptake of the tool may have related to one of these areas and it may be due to a lack of perceived new knowledge or value placed

on evaluation tools. These factors are also worth considering when considering the consequences of the tool.

The small sample size is a marked limitation of this study however despite small numbers significant correlations were found when examining the reliability of the toolkit and its relation to other variables. As the data was collected in local units I had to take into account how training actually takes place and use this naturalistic setting to investigate reliability as far as possible. Due to these naturalistic settings it was not possible to use Generalisability theory because the data could not be collected in such a way that different sources of variation could be identified. Instead I used Classical test theory in order to examine the data; it was possible to examine test-retest reliability and interrater reliability.

The test-retest reliability examines how stable the results are over time. The toolkit showed good test-retest reliability with a Spearmans Rho of 0.751 ($p<0.001$). One would not expect the reliability to be a hundred percent between the two sessions as these two sessions would not have been the same. For instance each session would have contained different cases with differing levels of difficulty which classical test theory does not allow us to account for. Other aspects between the two lists such as differing nurses assisting or different time pressures would also have existed. Even given these factors the trainer's ability to teach appears to be consistently rated by the DOTS over time. If the toolkit were to be used for summative use any variation over time would be deemed negative but given the fact that the toolkit is intended for formative use then some variation can give useful insights. Harlen and James (1997) argue that one might expect some differences over time in formative evaluation or assessment particularly when the test situation is changeable. For instance if a trainer were to be evaluated highly on a list of relatively straightforward procedures but is evaluated more poorly on a list of difficult cases that becomes time pressured then this is important information that can allow the trainer to reflect on the ways in which they trained differently. This is not to say that the trainer should not change their teaching if the list overruns but this needs to be communicated appropriately to the trainee.

There was a moderate inter-rater reliability between trainers and trainees when they completed the DOTS with a Spearman Rho of 0.516. Discrepancy between trainee and self-evaluations has been seen previously in other studies (Claridge, Calland et al. 2003) where 61% of teachers scored themselves significantly differently from their learners.

One might hypothesise that this has occurred for several reasons; one argument as discussed above is that learners typically rate teachers highly leading to discrepancy between the two sets of scores. A different argument may be that the teachers are more self-critical of themselves which may have been emphasised by the fact that these were self-selected group of teachers who are interested in their teaching and therefore more self-critical. Another explanation for this difference in scores may be explained by the differences seen when examining the internal structure of the tool, clearly there are certain items that seem to be evaluated differently by trainers and trainees, this may relate to the fact that trainees and trainers are both actively involved in the training but from different viewpoints which may make it more or less easy to measure each attribute. This is a disadvantage of using the same tool items for both trainees and trainers in that all the items may not be as easily measurable by both groups. This may also act as an advantage as it might promote reflection in trainers in areas that trainees have evaluated differently; however if the discrepancy is always that the trainee has scored the trainer more highly then one could argue that all this discrepancy may lead to is an improvement in the trainer's confidence and may be of limited use in improving teaching. Before discounting self-assessment on this basis however it is important to bear in mind again that this was a self-selected group of trainers and therefore this may not always be the case. In previous studies those trainers that did not self-evaluate were evaluated less well by their trainees (Claridge, Calland et al. 2003). As previously discussed even within this small sample size there was a suggestion from the free-text comments that completing the self-assessment had led to a change in practice for one trainer.

Using the 'Training the trainer' courses to gain peer evaluations was not as ecologically valid as collecting the data within local units as I eventually want peers to use the tool within local units. This means that I have not been able to see if there are practical difficulties with gaining peer evaluations, for instance whether there is enough room for a peer to also be in the endoscopy room. At this stage however I felt the priority was to evaluate the internal structure and reliability and it was advantageous to gain multiple peer assessments in order to gain a fuller assessment of the psychometric properties, for instance this allowed factor analysis to be performed. It also used similar peers, other training endoscopists to perform the evaluations.

Gaining multiple peer assessments also meant that Generalisability theory could be used to examine the reliability of the tool, meaning that one could more closely examine for sources of variance. A disadvantage of using the 'Training the trainer' courses was that the study design was still constrained by the way in which the course is run. One such disadvantage was that each trainer only has the opportunity to train on one case therefore it was not possible to examine for the effect that different cases have on training ability.

Initially when carrying out the Generalisability analysis I hypothesised what the main sources of variance were that caused variation in scores. These were differences in trainers' ability to teach (the object of measurement), variation in how peers evaluate, as those who complete evaluations have a natural tendency to either leniency or stringency within their evaluations. As well as their natural tendencies in how they evaluate peers will interact with different trainers differently aside from that trainer's actual training ability. This may be because that trainer trains in a similar way to themselves which they then favour or perhaps does something in a way they do not like but does not actually reflect poor training skill.

When these sources of variance were inputted into the generalisabilty analysis trainer variance i.e. true differences in the trainer's ability to teach accounted for 44% of the variance seen within the observed score. The next greatest source of variance was due to peers, which explained 34% of the variance, followed by trainer: peer interaction which accounted for 22% of the variance. This was reassuring that differences in trainers accounted for the greatest amount of variance within scores. It was then possible to calculate G co-efficients, which are similar to Reliability co-efficients, for different number of raters. One peer rater had a G-coefficient of 0.44 and three raters were required to gain a G co-efficient of 0.7.

The G coefficient gives an idea of how reliably the scale reflects variance in the object of measurement but it does not tell test constructors or reviewers what this means in terms of an individual's score on the scale. It is therefore possible to calculate the standard error of measurement of the scale and the 95% confidence interval. If only one reviewer was used then the 95% confidence interval was very large with the true score likely to be found somewhere in the range of 13 points above or below the observed score. For instance if a trainer received a score of 64 from one peer evaluator then it is possible to say with 95% certainty that their true score lies somewhere in the

range of 51 to 77; this range is clearly quite large and means that they could either be in the top or bottom quartile of trainers. Such a spread of scores is likely to be unacceptable. As the number of peers increases so does the reliability of the DOTS and as it becomes more reliable the range between the top and bottom confidence interval becomes smaller. For instance the range for 95% confidence intervals with five reviewers was 7 points above and below the observed score.

A second generalizability analysis was performed to explore for other sources of variance. The model that was produced was of poor fit and suggested the initial analysis had given the best explanation for sources of variance. This may however be due to the manner in which the data was collected; for instance in this second analysis course explained 22% of the variance in scores; however as the data was nested in course no degrees of freedom were reported. There was however some evidence that course may have affected the reliability; when the data was examined separately by course there was wide variation in the confidence intervals between courses. A one way ANOVA examining course further suggested a significant difference between courses. There may be several explanations for this. One of the reasons may be that the tool itself may have been used differently on different courses. I attended one of the courses, introduced the project and administered the tool to peers; when a separate generalisability analysis was run for this course it demonstrated a very high reliability. On other courses the tool was administered either by one of the faculty teaching on the course or a member of the endoscopy administration staff at that unit. The degree of explanation and relative importance that they placed on the tool may have affected how it was completed and the attention paid to completing the tool. My presence may also have altered the degree of missing data and may have added a further variable to the reliability, however clearly in the future the aim would be for the tool to be administered locally.

An alternative reason that courses accounted for so much variation within scores may be due to factors directly related to the course itself. All courses are generally taught to the same format as suggested by JAG (personal communication JRB) and therefore one would not necessarily expect courses to be a source of variation. This may suggest that there is something different about the courses aside from the curriculum, which suggests perhaps that there are differences within the culture of the courses. It may be that on some courses those attending the course form more of a relationship with each

other. These relationships may encourage participants to evaluate their peers more or less harshly, in that they feel more comfortable and are therefore more honest in their evaluations or that they 'like' others on the course and therefore evaluate them highly regardless of actual performance. It is also possible on some courses that some course participants may have already been known to each other as sometimes those from the same department elect to go on the same course (personal observation). On those courses on which this occurred this previous relationship with others may have impacted on the evaluations given.

The other factor may be something to do with the culture of teaching created by the faculty. The course clearly focuses on training and therefore an ethos of a desire to improve in teaching should be cultivated. One could argue that if such a culture were to exist on the courses then there should be a desire to evaluate others on the course fairly as a desire to help them improve should exist. If such a culture does not exist then this may be reflected in less accuracy within the tool. As there was no crossover between the courses it is not possible to examine for this using a generalisability study but one could perhaps examine different behaviours on the course using an observational study if one wanted to examine this further.

The other limitation with this study is that it was not possible to examine the variation caused by different cases. Different cases are likely to differ in terms of their difficulty and therefore necessitate different trainer skills but also will not be equally difficult for all trainers in that individual trainers are more likely to feel more or less comfortable with different problems. The difficulty of case is likely to affect the trainers score but as each trainer taught on only one unique case, case was nested in trainer and it was not possible to consider variation due to case. This is worth acknowledging as above I have discussed the numbers of peers required to watch one case in order to provide adequate reliability of the DOTS. Clearly in real units having several peers observe a single case may not be practical, additionally once the effect of course had been inputted into the analysis the number of peer reviewers required to gain adequate reliability became impractical. It may be that an alternative to increasing the number of reviewers required may be to increase the number of cases observed; this would be an alternative method of increasing the reliability of the tool but as trainers only taught on one case it was not possible to examine for case within this design.

9.4 Relation to Other Variables

In order to compare the toolkit to a previously validated scale when completing the LETS participants were asked to complete the Clinical Teaching Effectiveness Inventory (CTEI) (Copeland and Hewson 2000). When the results of the CTEI were compared to the results of the LETS there was a high degree of correlation between the two tools (Spearman Rho = 0.868) that was statistically significant ($p<0.05$). This correlation suggests that these two tools are measuring the same construct, that of good clinical teaching. The CTEI was used for comparison as it has previously been tested with large sample sizes and evidence is provided to support its validity (Copeland and Hewson 2000). It has also been shown to have good evidence of reliability with a G co-efficient of 0.742 with just one rater (ibid) which was important given that only one CTEI was completed by a trainee for each trainer. There are no reports that the CTEI has been used to evaluate endoscopy trainers but it has been used to evaluate a variety of different kinds of clinical teachers (ibid) across different institutions in different countries (Hemstokroos and Vleuten 2005). The CTEI was also chosen because it is short in length (15 items). The very high degree of correlation between the two tools is evidence that the LETS measures attributes of good teaching as this is what the CTEI also purports to measure.

It is important to note that the reliability of the CTEI is not always consistent across different institutions. When trialled in Dutch medical schools the reliability was found to be lower with a G co-efficient of only 0.4 with one reviewer(Hemstokroos and Vleuten 2005) this demonstrates that the validity of the tool is specific to its test circumstances and therefore although I have used the CTEI as it was a previously validated tool the results must be interpreted with caution. Additionally I have also used both trainee and trainer evaluations to calculate the degree of correlation between the two tools. The CTEI has not previously been used as a self-evaluation tool therefore there is no evidence about how it performs within these test circumstances. However if the trainer data is excluded there is still a strong correlation between the two tools although this does not reach statistical significance due to the small sample size. Other studies have used previously validated tools in a similar way to validate new tools. Steiner(Steiner, Franc-Law et al. 2000) compared a new score to evaluate clinical teaching performance within the Emergency Room (ER) with a scale developed by Irby(Irby 1988) that evaluated teaching within inpatient and outpatient setting. Steiner et al (2000) used Irby's tool as they felt that teaching within the ER was similar

to teaching in outpatients. They did not re-validate Irby's tool initially but felt that a correlation between the two tools was still evidence of their new tool's validity.

The other alternative would have been to compare the results of the toolkit to the pre-existing tool used by JAG (JAG 2012); however this tool has no previous psychometric assessment of its measurement properties. Indeed there are no previous tools that have been empirically tested within endoscopy and therefore I felt that comparing the toolkit to a tool that has at least been validated within a number of different fields would be a more appropriate comparison and be stronger evidence for the toolkit's validity.

The presumption is that the as the LETS and the CTEI correlate well then they must both be measuring the same construct, that of effective clinical teaching. It is however hypothetically possible that they could both be tapping into a different construct. This is unlikely given the work that has gone into ensuring the content validity of both tools but cannot be definitively proved. The other option therefore would be to compare the toolkit to a different variable; the variable that several evaluation tools have been compared to is student performance. This comparison relies on the hypothesis that good teaching ultimately leads to an improvement in learning and student outcomes. Two studies within the same institution compared student evaluations of teachers to student performance in examinations (GriffithIII, Wilson et al. 1997, Blue, III et al. 1999). Griffith III et al (1997) compared the evaluations of general internal medicine attending to student scores on both a written and a practical examination. They found that students who were attached to the 'best' teachers (those who received the top 20% of evaluations) had statistically significant better scores on their written examination whilst those who were taught by the 'worst' teachers performed less well in a practical examination. The authors hypothesise that this was likely because those with the worst teachers did not have an effective role model from which to learn the skills required in the practical exam such as communication skills. Blue et al (1999) also found that those with the best surgical teachers also did better in some parts of the surgical exam although no difference was seen in other parts of the exam. I felt that comparing the LETS or DOTS to trainee outcomes as a measure of whether they accurately measure teaching effectiveness was not feasible. In Chapter 1I discussed that using learning outcomes as a measure of teaching effectiveness would be difficult as the regularity of measurable learning outcomes within endoscopy is sporadic and

there can be long time periods before the next milestone is reached. Trainees are encouraged to collect formative DOPS but these are often carried out by their trainers and therefore would be an inappropriate marker of teaching effectiveness.

Using the CTEI as a comparative tool may not have been without limitations however I feel that it is still the most appropriate choice. The fact that the study was designed in order that relationship to other variables was considered remains a strength of this study particularly as reviews of other evaluation tools have shown that this is a source of validity evidence which is often not considered; Beckman et al (2005) argue that it is 'a powerful yet underutilized source of evidence(pg.1162)'.

9.5 Consequences

Evidence of the consequences of the toolkit refers to providing evidence of the impact that the toolkit has on those whom are evaluated. This project has focused on the development of the toolkit and feedback has not been given to trainers; this area represents further work that needs to be carried out to develop the toolkit.

Prior to giving feedback to trainers it would be necessary to consider how that feedback should be presented. Considering the surgical tools discussed in Chapter 2 the methods by which the results of the evaluation was given as feedback to the teachers varied from study to study. An evaluation tool intended for summative use(Cohen, MacRae et al. 1996) only provided a mean score for trainers, which the authors describe as a 'teaching effectiveness score'. This may be useful for summative use in order to rank or highlight those teachers that excel but does not inform teachers how to improve. Other surgical evaluation tools provided a mean score for every item (Downing, English et al. 1983, Tortolani, Risucci et al. 1991). This provides the trainer with the maximum detail from the tool possible, which one could argue is beneficial for formative use as it would give the fullest picture of a trainer's strengths and weaknesses, however in considering the content validity of the tool the strength of basing a tool on a theoretical model was discussed. Bowden and Marton (2000) argue that if a tool has theoretical underpinnings then it is most likely to bring about effective change in a teacher's behaviour. Just listing a mean score for every item does not demonstrate any theory of teaching or learning. Additionally the amount of detail provided might make it less meaningful to the teacher as they would be unable to retain the information about how they had performed on every item. As a result of this

it may be useful to group theoretically related items and feedback domain scores as well as individual item scores.

There are several options for how scores could be combined or grouped in order to try and provide the most meaningful feedback. One method would be to use one of the theories or models discussed in Section 9.1.2. Using one of the established theories however would be inadequate as neither Maslow's theory of learning nor Collins' Cognitive Apprenticeship model fully explain the process of teaching and learning of endoscopy and all the items are not categorised using a single theory. A further option would be to use Wells' model of effective endoscopy teaching; however as all the attributes on the DOTS only relate to two domains then this would lead to limited detail. Combining the results of the two tools would provide a fuller representation of the model but may be confusing given the different frequencies of the tools.

Another option to feedback the results of the DOTS would be to use the results of the factor analysis which emerged when the DOTS was completed by peers (Chapter 7); the four factors are shown in Table 9-3. If one were to consider feeding back the results of the DOTS within these domains one would need to bear in mind that this factor structure was seen when the DOTS was completed by peers only therefore one would need to collect more trainee and self-evaluation data to see if the same domains existed. Confirmatory factor analysis could be used specifying the same factors as found for the peer's data to see if the same factors existed within the DOTS when it is used by trainees and trainers. Additionally one would need to consider adding more items to factors two and three as factors with only two items are statistically weak (Costello and Osbourne 2005) and may not be found if more data was collected and analysed.

Alongside their own feedback the subjects of the surgical evaluation tools were also often given an idea of how they compared to their colleagues who had also been evaluated(Cohen, MacRae et al. 1996, Iwaszkiewicz, DaRosa et al. 2008). This enables teachers to gain meaning of the score they have been given relative to their colleagues. One of the advantages of this is that it may motivate trainers to improve in order to 'beat' their colleagues. This may mean that improvements to teaching would be made away from a true desire to be an effective teacher and this competitive element appears at odds with the formative purpose of this tool. It may however make their score more meaningful than just a number as it gives an idea as to whether they are

excelling as one of the best trainers or one of the worst particularly given that we are aware that using the toolkit trainees tend to score their trainers highly. In this situation it may give more meaning to high scores as it may make those that have received a 'good' score appreciate that with respect to their peers they could still improve and are not seen as one of the 'best' trainers.

The other consideration is not only what information should be given as feedback to trainers but also the method used to deliver the feedback. One of the dangers of giving feedback to trainers is that it may have a negative effect on those who are evaluated poorly as was seen by the study performed by Litzelman et al (1998). In their study feedback was given in the form of a written report of the participant's scores, alongside an 'individualised report' which emphasised the categories in which the subject's had performed less well compared to peers. Subjects were also made aware of a 'teaching effectiveness service' which they could access at their own will although no subjects sought out this service. In contrast other studies have found that the lowest scoring teachers initially improved the most compared to those that scored more highly (Cohan, Dunnick et al. 1995, Maker, Lewis et al. 2006). The reason for this difference between studies may be that in the latter two studies the results of the evaluations were discussed with the teacher in person rather than just receiving a paper copy of the results. In both studies this verbal discussion of the feedback occurred with the program or faculty director. Some of the suggested reasons why improvement was not seen in the lower scoring faculty in the study conducted by Litzelman et al (1998) were that the lower scoring faculty were already trying their hardest and were discouraged by the poor feedback or did not have the necessary skills to bring about change in their teaching practices. A discussion of their teaching feedback may therefore have been useful. This discussion could have been used to continue to encourage teachers and also make practical suggestions for change. The fact that a senior figure was involved in the feedback may have also given credence to the evaluation process and made teaching staff aware of the importance that senior figures placed on teaching within the department. This would be worthwhile considering when giving the feedback to endoscopy trainers. Each endoscopy department has a training lead whom could be involved in such feedback; this may also be helpful in integrating the feedback from the two tools and ensuring that equal importance is given to both the technical and interpersonal attributes. Brinko (1993) also argues that feedback is more likely to bring

about change when it is mediated by a consultant. She argues that the consultant should act as a facilitator and should help the subject identify problems and set goals rather than assume the role of expert. Given this knowledge it may be that if such discussions were to take place then those leading the discussion may require training. The above is also highlighted by Centra (1993) who states that change can only result from evaluation if the trainer or teacher knows how to change. Just giving a trainer the results of their evaluation may not fulfil this condition, meaning that future development is less likely.

One of the practical issues regarding giving feedback to endoscopy trainers is that of anonymity. The standard in most studies of evaluation is that learner respondents remain anonymous to the subjects so that it is not possible for subjects to determine what score each learner has given to each trainer. The reason for this is that it is felt to promote more honesty and avoid falsely high ratings within the feedback. This is supported by a study that compared open and anonymous evaluations which found that the mean score on open evaluations was statistically higher to that on anonymous evaluations (Afonso, Cardozo et al. 2005). This leads to practical difficulty when evaluating endoscopy trainers as they may have only one or two trainees attached at any one time and therefore maintaining anonymity may be a challenge. A counter argument to this is that if the ultimate goal is to improve teaching then open honest evaluations should be encouraged (Goldstein 2005, Guerrasio and Weissberg 2012). All endoscopy trainees are adult learners who often outwith endoscopy supervise other juniors and should be used to giving constructive feedback, therefore there is an argument that this is no different to giving feedback to their trainers. However, as a trainee may wish to apply for a job with that trainer in the future then giving such feedback may be difficult. The data collected in this study was all collected with the assurance of anonymity, if one were to consider using open evaluations in the future further work on the reliability would need to be performed.

Clearly further work is needed to understand how to present the data from the toolkit to trainers and what methods of feedback would be most acceptable and most effective in bringing about change. The other area in which further work is required is considering how to integrate self and peer evaluations alongside trainee evaluations and to consider how often the evaluations need to occur. The current system used by JAG to collect feedback is that their feedback tool appears after every list recorded on

the JETS website (JAG 2012). This means that feedback from the trainee occurs regularly as there is a system by which it is easily collected. In terms of self-evaluation, the main evidence of its purpose is that trainers feel that it helps promote reflection on their teaching. There is a danger however that if I were to ask trainers to complete a self-assessment after every list then this becomes 'reflection overload' and becomes a paper exercise without any true reflection occurring. I was unable to find any literature which suggests the ideal frequency of self-assessment and therefore this would be an area for further exploration. Similarly it would not be possible to gain a peer evaluation of every list. In this study I struggled to engage trainers and trainees to participate in the study, this would have been even more challenging were I to have also required a peer evaluation therefore the frequency of any such evaluation would need to be balanced with the time taken by a peer to observe a session.

The other concern with any toolkit is that is it enough to bring about change? Further studies could explore trainers' perceptions of feedback from the tool to discover how useful it had been in improving their teaching and whether once feedback was delivered there was any change in future ratings. It is a limitation of this study that we did not explore the utility of the toolkit in aiding trainers self development. Clearly this would be more helpful once trainers are receiving feedback from the toolkit however provisional work could have been commenced to start exploring this area. It could have been linked to the self evaluation and asked whether they thought the toolkit had utility. Further work to develop the toolkit could therefore concentrate on the process of delivering feedback to trainers. This could include a qualitative study to look at trainer's attitudes to evaluation which may help explore why uptake was so poor in this study. The different methods by which feedback is given could also be explored by providing feedback to trainers using a variety of methods. One could then either interview trainers to explore their reactions to the different methods or one could continue to gather subsequent evaluations and investigate whether the manner in which feedback is given affects future performance.

9.5.1 Future directions

Much of the future work to further develop this project centres on evaluating the consequences of the tool as discussed above, hence the reason for discussing future directions in this section. In addition to considering the consequences of the tool one

of the other pieces of work, which may contribute to this work, is a study detailing how endoscopy training is delivered in different units. As discussed previously, in some units a trainee is attached to a single trainer however in other units trainees sign up to individual lists in order to try and maximise learning opportunities around other clinical commitments such as on-call shifts. This may impact on the use of the DOTS and LETS, as discussed in section 9.1.2.1, and therefore it may be helpful to gain understanding of what is currently happening with regards to training in individual units.

In order to explore the consequences further data collection is required. This will also help gain further assessment of the internal structure of the tools and discussions are planned to take place with JAG in order to develop this further with the possibility that it may be adopted onto the JETS eportfolio system which will aid with data collection. Support for this adoption would come from the fact this project has shown reasonable validity and reliability for the toolkit, in particular compared with the evidence for the existing JETS tool for trainees. Collection of this data would enable further assessment of the internal structure of the tool including confirmatory factor analysis and reliability. Further study would then centre on assessing the consequences of the toolkit as discussed above. This would include a qualitative element focusing on whether trainers found the feedback from the toolkit useful but also assessing whether feedback from the toolkit improved scores. The methods by which this feedback is given would also be explored including how the data is presented and by whom, whether this is through the JETS website, whether comparison to other scores within the department make a difference and the involvement of training leads in discussing the results of evaluation. This would help explore the consequences of the toolkit but also gain further evidence for all the components of validity and further strengthen the tool. These suggestions for further work are shown in figure 9.2.

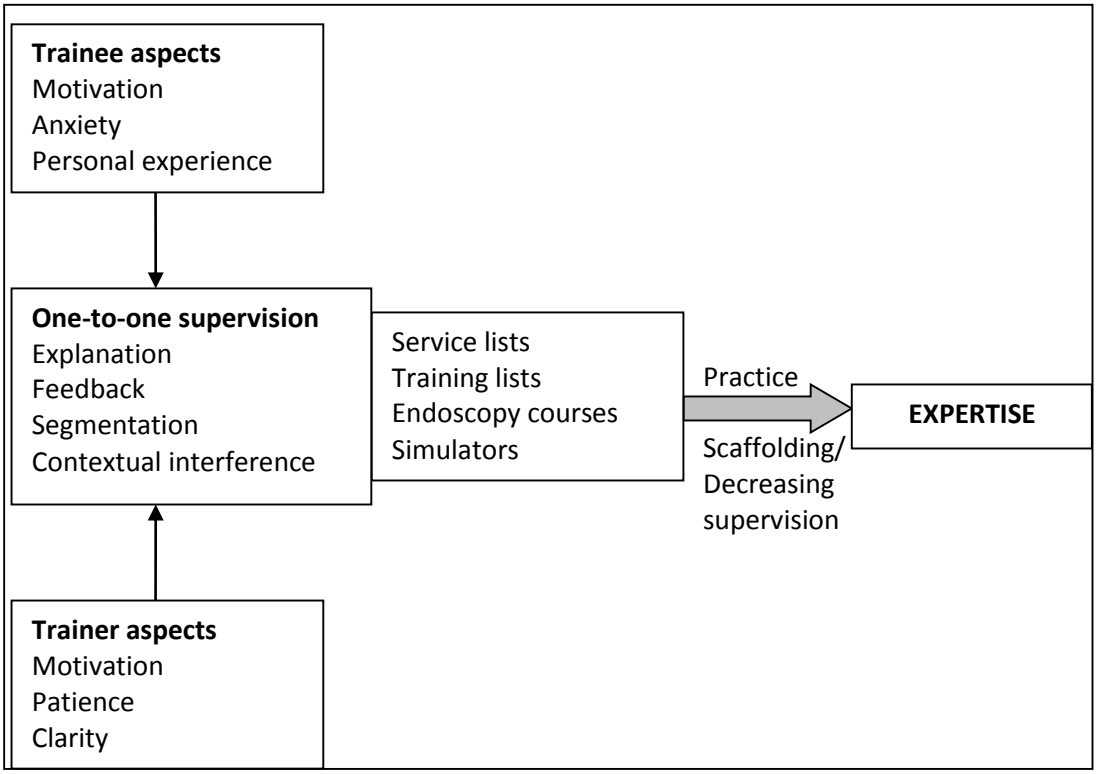
Figure 9-2 Flowchart depicting possible areas for future work from this study

9.6 Conclusion

The above discussion considers the strength and limitations of both the toolkit and the methods I have used to derive and test it. The discussion has focused on the trainer but it is important to note that training will also be affected by factors relating to the trainee. Not every trainee will respond to the same trainer attributes (Sutkin, Wagner et al. 2008) and even excellent trainers will not match all trainees' learning styles (Stern, Williams et al. 2000). This interplay between trainer and trainee is difficult to capture in an evaluation. This was highlighted in Thuraisingam et al's (2006) work on the endoscopy learning experience. In this study the opinions of trainees as to what constitutes a good training experience was sought. From the themes that emerged Thuraisingam (ibid) proposed a model for the endoscopy learning experience shown in Figure 9-3. This model suggests that the process of one-to-one supervision is at the centre of the endoscopy learning experience and highlights that as a trainee moves towards expertise supervision changes. The model incorporated both trainer and trainee attributes that impact on the training experience. In attempting to evaluate the endoscopy trainer it is therefore important to acknowledge that such an evaluation will

be affected by the trainee. Not only in respect to the fact that different trainees will have different opinions as to what is an effective trainer but also that how the trainer trains is affected by the differing attributes of a trainee. Thuraisingam's model also highlights that the level of competency of the trainee affects the training process. I have attempted to ensure therefore that the items are suitable for all levels of trainee but it was not possible with the small data set to compare different grades of trainee when trialling the toolkit in local units.

Figure 9-3. Schematic model of the endoscopy learning experience suggested by Thuraisingam et al. (2006)



Throughout this thesis I have recurrently emphasised that the aim of the toolkit was to be formative, this relates back to the ultimate aim of the project was, not to try and assess the standard of endoscopy training, but to try and improve endoscopy training. Following a review of the literature it was hoped that a toolkit may be a method to provide feedback which could subsequent lead to an improvement in training. Given this project has subsequently focused on deciding what to measure and ensuring they are measured in a 'valid' way. Now the toolkit has been created one does need to consider its potential uses; despite being created for formative use, could it be also used as a summative tool? As discussed in section 2.6.1 the DOPS tool for trainees is used in both a formative manner to give feedback and as a summative tool to assess whether competence has been reached. One could argue that in terms of teaching on

one of the JAG approved training courses or the Training the Trainer course only excellent teachers should be faculty although currently no method exists to assess their 'excellence'. The GMC state that teaching excellence should be recognised and rewarded (GMC 2011) and although they state a multi-faceted approach should be used to assess further detail is not provided. Clearly, as discussed in section 9.3.2, the toolkit has not yet shown sufficient reliability in order for high stakes decisions to be made as this requires a reliability co-efficient of greater than 0.9 (Downing 2004) however if the toolkit were subsequently to show such reliability it perhaps could be used as part of a method to select faculty for courses. The dilemma also arises as to whether the toolkit should or could be used in a summative manner in order to make decisions about whether poorly performing trainers should have their training responsibilities removed. The GMC (2011) recognise that not everyone is good at teaching and one could argue that if a trainer's skills are poor then one has to consider the impact on a trainee's progression and learning and their future competence as an endoscopist. If the tool were to be used in such a manner, again one would need to be confident that it was able to discriminate between teaching ability very highly. This also does feel counterintuitive to the formative intent of the tool but poor results should not be ignored. This further strengthens the argument for the involvement of a trust's endoscopy lead in delivering and discussing feedback as then any trainer who is doing poorly would be given the tools to improve. If there were a failure to improve hopefully this would lead to further discussion between the trainer and training lead as to whether cessation of training should be considered.

Both Wells (2010) and the Delphi focused on the concept that the toolkit must contain measurable items, this in itself ultimately led to the exclusion of some items. For instance one of the items that was excluded was

The trainer dealt with any lack of insight in the trainee

This was excluded by the Delphi panel as it was felt that it would not be possible for the trainee to evaluate such an attribute. This remains an important attribute for the trainer to possess though; endoscopy is a potentially hazardous procedure with serious complications, therefore it is important that a trainee recognises their own limitations and seeks help appropriately. If a trainee does not recognise such limitations it is important that this is addressed by the trainer; the fact that it has not been included in

the toolkit does not mean that this attribute is less important than those that were included, just less measurable. Wells (2010) suggested that alongside the toolkit a handbook should be introduced which could also include those attributes that were not suitable for inclusion in the toolkit. This would mean that these attributes would still be recognised as important and not lost. Such a handbook could be used to complement the evaluation toolkit. Sarker et al (2005) discuss in their development of a toolkit to evaluate technical teaching skills within the operating room that their aim for this tool as well as to provide formative feedback is that they later hope that it will act as a guide to how such teaching should occur. If a similar process were to occur within endoscopy training then a handbook would be useful to ensure that those items that were more difficult, or not possible, to measure would still be considered.

In summary

- The attributes described by Wells (2010) have been successfully used to create an evaluation toolkit for endoscopy trainers. The use of these attributes and their selection through the Delphi process has provided strong evidence for the content validity of the toolkit.
- The resulting DOTS and LETS reflect Wells' (2010) model of endoscopy teaching, Collins'(1989) cognitive apprenticeship model and Maslow's (1970) theory of learning, however none of these theories can be used to fully explain all the attributes on the toolkit
- Those attributes that could be described as interpersonal were all included on the LETS; this means that if only the DOTS were used these interpersonal attributes would not be captured therefore trainers in those units who do not allocate trainees to a particular trainer may not be evaluated on these skills
- On trialling the tools a high internal consistency was demonstrated with all user groups; however at an individual item level there were differences in the degree of correlation when the tool was used by different groups. This suggests that the items are not judged in the same way by all the different groups
- Peer data also enabled internal consistency to be examined further using factor analysis which suggested that although there was high internal consistency there were sub-domains within the construct of good endoscopy training

- Peer evaluation was reliable when the tool was used on training courses with 3 peers required for a G co-efficient of 0.7
- When trialled in local units the DOTS showed acceptable test-retest reliability. The inter-rater reliability between trainees and trainers using the tool as a self-evaluation was moderate; this has been seen previously in other studies
- There was a strong correlation between the LETS and the CTEI demonstrating evidence for relation to other variables
- It was difficult to engage trainers within local units to trial the tool; reasons for why this might have been require further exploration. In addition further work is required to investigate how the results of the tool should be given as feedback to trainers and the impact of such feedback
- As the tool does not include all the attributes of a high-quality endoscopy trainer, a handbook of attributes may also be useful to accompany the toolkit

Chapter 10. My Development as a Researcher

This dissertation presents the research that I undertook to try and produce a validated toolkit that could be utilised to give feedback to endoscopy trainers regarding their training of endoscopy trainees. The discussion in the preceding chapter highlights some of the strengths and weaknesses of both the tool itself and the methodologies used to derive and test it. In this chapter I want to reflect on my development as a researcher during this project and my interaction with both the data and the methods used.

Prior to commencing this MD I had just completed my rotations in Core Medical Training; I had received no formal education in teaching or learning and had not yet started to learn endoscopy. Whilst completing this research project I have learnt how to perform diagnostic upper GI endoscopy and have been signed off as competent. I have also completed the Certificate of Clinical Education at Newcastle University which contained a module on learning theories.

Throughout this study I was a novice endoscopist, and at the beginning had no real experience of endoscopy training. I had however throughout my prior medical training had experience of being taught practical skills, albeit less complex ones. My lack of experience may have had advantages to the project, as I had no prior experiences or strong opinions that could have influenced the results of the study. One could argue that I was analysing the topic through 'neutral eyes'. An experienced endoscopist, however, may have had a wealth of experience to draw on which may have altered the way in which they viewed and analysed the data and more practical experience of how such a toolkit might be used in different units. The research team included experienced endoscopists. This meant that in discussions they were able to discuss the data in relation to their different experiences, which I believe has meant my novice status has not unduly affected the outcomes of this study.

One of the strengths of this study was the contribution that Wells' (2010) work has made to the content of the toolkit. This project has essentially been a continuation of his work and this had been the intended outcome from the start. The supervisory team were the same for both projects and this from a personal perspective was both a strength and a weakness. It was helpful in that my supervisors had a very good understanding of his work and in-depth knowledge into how the attributes were

derived. It was also useful as I was able to access Wells' N-Vivo files enabling me to access excerpts of the interviews that supported each attribute. As previously discussed in Chapter 3 this was beneficial in gaining greater understanding of the attributes. It was also helpful when performing cognitive interviewing (Chapter 3) if there was doubt to the correct interpretation of an attribute. Access to this information has strengthened the content validity of the toolkit as the attributes continue to mirror what was said in the original interviews. Continuing work started by another individual however has also had its challenges. At the beginning there was a sense of playing catch up with regards to the knowledge but also in having to make sure that the decisions that were made reflected my decisions rather than decisions that were made prior to my commencement on the project. This was something that developed as the project progressed and I was able to take more ownership of the data.

During my time in research I have learnt that I am undoubtedly a positivist. A positivist is someone who believes that knowledge is objective and can be measured accurately from observation and is either correct or incorrect (Cohen, Manion et al. 2007). As a doctor I have been educated from a scientific background and then worked in a world of hard facts. Accepting that all knowledge may not be such a case of either 'right' or 'wrong' has been a personal struggle throughout this project. There have not always been clear results that have suggested that something is right and wrong. The result and the surrounding literature have therefore needed to be considered to try and determine what the 'best' choice was. I often discovered that there was not always a right choice and that 'best' was also not often so clear cut. One example of this has been using the Delphi both in determining how to conduct a Delphi study and marrying the quantitative and qualitative results. The Delphi process is well described in the literature but there is no 'recipe' as to how it should be conducted. A myriad of different topics have been considered using a Delphi process but all used slightly different methodology. This was my first experience of research and was somewhat overwhelming; particularly in relation to the handling of the results and trying to interpret the statistics as well as the free-text comments. In making decisions about the statements there did not always appear to be an absolutely right answer. Using the rules suggested by Yeates et al (2008) was helpful in trying to make this a systematic process, considering the general themes that arose from the panel was also useful. Even when trialling the tool there was not always an absolute right or wrong with

regard to the statistical analysis. Statistics such as the factor analysis require a level of interpretation as they do not present an absolute answer, rather one must interpret the strength of the model based on issues such as deciding how to extract factors, crossloading and correlations. I initially found this challenging as I wanted the factors to either exist or not exist. Working on this project has enabled me to have this insight and to understand that my initial reaction is to be frustrated by those things that do not fit my positivist framework. It has enabled me to be able to consider things more carefully and accept that I have to try and make a balanced opinion on the evidence available rather than just being told whether an answer or method is right or wrong. I have also accepted there are many ways that things can be done and one method is not necessarily better than another however what is important is to make a reasoned choice.

Dealing and accepting the limitations of the study have also been one of the challenges that I have faced through this project. One of the 'mistakes' I made was in calculating percentage agreement in round one of the Delphi. On initial analysis I inadvertently also included the amount of missing data within my analysis. I only realised this error after sending out the data in round 2. This meant that in fact the cut-off for agreement was 77% not 70%; this was an error and taught me the importance of triple checking calculations. It was a useful learning curve in considering how to deal with it and considering the strengths and weaknesses of the various options. In the end I decided to raise the consensus level to 77%; this does alter the results of the Delphi and highlighted that this is a potential weakness of the methodology as this could be used to manipulate the results.

The other major limitation was that despite best efforts the data collected in local units was very small. Despite this small sample size most of the results still reached statistical significance which was reassuring but gaining more data would have enabled me to consider the LETS in greater detail. This is very much a limitation of the study but in writing the discussion I realised that it was important to consider why this might have been and potentially means that further consideration has to be given to how useful a process trainers think this is. I realised that rather than being disappointed by the result it was important to address it and consider why it had occurred.

At the beginning of the process of conducting this research project I struggled with what should be included in the literature review and I felt that the lines were more

blurred compared to a more scientific MD but as I wrote the discussion I realised that this helped pull together parts of the literature review. In particular considering the toolkit in relation to existing theory helped develop my understanding of what I was trying to achieve by evaluation. Trying to capture the essence of a good teacher is a challenge (Sutkin, Wagner et al. 2008) and can mean different things to different people I have strived to develop a tool that reflects evidence for validity and has items that demonstrate statistical properties such as internal consistency and reliability. It is however important to realise that whilst I can strive to create a 'valid' and 'reliable' evaluation tool quality remains a relative concept and is only of value if those that use it believe in it. I feel that in accepting this whilst also balancing it with the psychometric properties has been the most important process of this project.

10.1 Presentations arising from this research

L Macdougall, S Corbett, C Wells, M Welfare J.R. Barton High quality teaching of endoscopy; devising a toolkit to evaluate the trainer' 15th Ottawa Conference on Assessment in Medicine (Association of Medical Education), oral presentation Kuala Lumpur March 2012

L Macdougall, S Corbett, M Welfare, C Wells, J.R. Barton. Evaluating endoscopy trainers, how reliable are peer evaluators? Poster presentation, BSG Glasgow 2013. Awarded poster of distinction and best poster in category

L Macdougall, S Corbett, M Welfare, C Wells, J.R. Barton. Devising a toolkit to evaluate the high-quality endoscopy trainer; a Delphi study. Poster presentation BSG Glasgow 2013

6 monthly oral presentations at Northumbria Healthcare Trust Research meetings detailing progress thus far and initial results 2010- 2012.

References

Acock, A. C. (2005). "Working With Missing Values." Journal of Marriage and Family **67**: 1012-1028.

Adshead, L., et al. (2006). "Introducing peer observation of teaching to GP teachers:a questionnaire study." Medical Teacher **28**(2): 68-73.

Afonso, N. M., et al. (2005). "Are Anonymous Evaluations a Better Assessment of Faculty Teaching Performance? A Comparative Analysis of Open and Anonymous Evaluation Processes." Family Medicine **37**(1): 43-47.

American Educational Research Association, et al. (1999). Standards for Educational and Psychological Testing. Washington DC.

Balfour, T. W. (2001). "Training for colonoscopy." Journal of the Royal Society of Medicine **94**: 160-161.

Barber, S. G. (1992). "Postgraduate teaching audit by peer review of videotape recordings." Medical Teacher **14**(2/3): 149 - 157.

Beckman, H. B. and R. M. Frankel (1994). "The Use of Videotape in Internal Medicine Training." Journal of General Internal Medicine **9**: 517-521.

Beckman, T., et al. (2005). "What is the Validity Evidence for Assessments of Clinical Teaching?" Journal of General Internal Medicine **20**: 1159-1164.

Beckman, T., et al. (2004). "A comparison of clinical teaching evaluations by resident and peer physicians." Medical Teacher **26**(4): 321-325.

Beckman, T., et al. (2003). "Evaluating an instrument for the peer review of inpatient teaching." Medical Teacher **25**(2): 131-135.

Beckman, T. J., et al. (2004). "How reliable are assessments of clinical teaching? A review of the published instruments." Journal of General Internal Medicine **19**(9): 971-977.

Berk, R. A. (2005). "Survey of 12 Strategies to Measure Teaching Effectiveness." International Journal of Teaching and Learning in Higher Education **17**(1): 48 - 62.

Billings-Gagliardi, S., et al. (2004). "Interpreting course evaluation results: insights from thinkaloud interviews with medical students." Medical Education **38**: 1061-1070.

Bing-You, R. G. and G. A. Stratos (1995). "Medical Students' Needs for Feedback From Residents During Clinical Clerkship Year." Teaching and Learning in Medicine **7**(3): 172 - 176.

Bisschops, R., et al. (2002). "A survey on gastorenterology training in Europe." Gut **50**: 724-729.

Blue, A. V., et al. (1999). "Surgical Teaching Quality Makes a Difference." American Journal of Surgery **177**: 86-89.

Boillat, M., et al. (2012). "Twelve tips for using the Objective Structured Teaching Exercise for faculty development." Medical Teacher **34**: 269-273.

Bowden, J. and F. Marton (2000). The University of Learning. London, Kogan Page Limited.

Bowles, C. J., et al. (2004). "A prospective study of colonoscopy practice in the UK today: are we adequately prepared for national colorectal cancer screening tomorrow?" Gut **53**(2): 277-283.

Bradley (2006) "The history of simulation in medical education and possible future directions" Medical Education **40**: 254-262

Brinko, K. T. (1993). "The Practice of Giving Feedback to Improve Teaching: What Is Effective?" The Journal of Higher Education **64**(5): 574 - 593.

BSCP (2011). "NHS Bowel Cancer Screening Program." Retrieved May 2011, from http://www.cancerscreening.nhs.uk/bowel/.

Burchell, B. and C. Marsh (1992). "The Effectr of questionnaire length on survey response." Quality and Quantity **26**(233-44).

Bye, A. M. E., et al. (2007). "A triangulated approach to the assessment of teaching in childhood epilepsy." Medical Teacher **29**: 255-257.

Campbell, S. and J. Cantrill (2000). "Prescribing Indicators for UK general practice: Delphi consultation study." British Medical Journal **321**(1-5): 1.

Carbone, E., et al. (2002). "Use of Cognitive Interview techniques in the development of nutrition surveys and interactive nutrition messages for low-income populations." Journal of the American Dietetic Association **102**(5): 690 - 696.

Chitsabesan, P., et al. (2006). "Describing clinical teachers' characteristics and behaviours using critical incidents and repertory grids." Medical Education **40**: 645-653.

Claridge, J. A., et al. (2003). "Comparing resident measurements to attending surgeon self-perceptions of surgical educators." The American Journal of Surgery **185**: 323-327.

Clark, L. A. and D. Watson (1995). "Constructing Validity: Basic Issues in Objective Scale Development." Psychological Assessment **7**(3): 309-319.

Clarke, D. M. (1999). "Measuring the quality of supervision and the training experience in psychiatry." Australian and New Zealand Journal of Psychiatry **33**: 248-252.

Clayton, M. (1997). "Delphi: a technique to harness expert opinion for critical decision-making tasks in education." Educational Psychology **17**(4): 373-386.

Cohan, R. H., et al. (1995). "Improvement of Faculty Teaching Performance: Efficacy of Resident Evaluations." Academic Radiology **3**(1): 63 - 67.

Cohen, J. (1960). "A coefficient of Agreement for Nominal scales." Educational and Psychological Assessment **20**: 37-46.

Cohen, J. (2008). Training and credentialing in gastrointestinal endoscopy. Advanced Digestive Endoscopy: Practice and Safety. P. Cotton, John Wiley and Sons**:** 289-362.

Cohen, L., et al. (2007). Research Methods in Education. Oxon, Routledge.

Cohen, P. (1980). "Effectiveness of student-rating feedback for improving college instruction: a meta-analysis of findings." Research in Higher Education **13**(4): 321 - 341.

Cohen, R., et al. (1996). "Teaching Effectiveness of surgeons." The American Journal of Surgery **171**: 612-614.

Collins, A., et al. (1989). Cognitive apprenticeship: Teaching the crafts of reading, writing and mathematics. Knowledge, learning and instruction: Essays is honor of Robert Glaser. L. B. Resnick. Hillsdale, Lawrence Erlbaum**:** 453-494.

Conigliaro, R. L. and T. D. Stratton (2010). "Assessing the quality of clinical teaching: a preliminary study." Medical Education **44**: 379-386.

Conrad, F. and J. Blair (1996). From impressions to data: increasing the objectivity of cognitive interviews. Proceedings of the Survey Research Methods Section of the American Statisical Association, American Statistical Association, Alexandria, VA.

Cook, D. and T. Beckman (2006). "Current concepts in validity and reliability for pyschometric instruments: theory and application." The American Journal of Medicine **199**(2): e6-e16.

Copeland, H. and M. G. Hewson (2000). "Developing and Testing an Instrument to Manage the Effectiveness of Clinical Teaching in an Academic Medical Center." Academic Medicine **75**(2): 161.

Cortina, J. M. (1993). "What is Coefficient Alpha? An examination of Theory and Applications." Journal of Applied Psychology **78**(1): 98-104.

Costello, A. B. and J. W. Osbourne (2005). "Best Practices in Exploratory Factor Analysis: Four Recommendations for Getting the Most From Your Analysis." Practical assessment, research and evaluation **10**(5): 3-9.

Cotton, P. (2008). Advanced Digestive Endoscopy: Practice and Safety, John Wiley and Sons.

Cox, S. S. and M. S. Swanson (2002). "Identification of teaching excellence in operating room and clinic settings." The American Journal of Surgery **183**: 251-255.

Crisp, J. and D. Pelletier (1997). "The Delphi Method?" Nursing Research **46**(2): 116-118.

Crossley, J., et al. (2002). "Generalisability: a key to unlock professional assessment." Medical Education **36**: 972-978.

Crossley, J., et al. (2002). "Assessing Health Professionals." Medical Education **36**: 800-804.

Crossley, J., et al. (2007). "I'm pickin' up good regressions':the governance of generalisability analyses." Medical Education **41**: 926-934.

Davis, D. A., et al. (2006). "Accuracy of physician self-assessmnet compared with observed measures of competence." Journal of the American Medical Association **296**(6): 1094 - 1102.

DeVon, H. A., et al. (2007). "A pyschometric toolbox for testing Validity and Reliability." Journal of Nursing Scholarship **39**(2): 155-164.

Dolmans, D. H. J. M., et al. (2004). "Providing physicians with feedback on how they supervise students during patient contacts." Medical Teacher **26**(5): 409-414.

Donnelly, M. B. and J. O. Woolliscroft (1989). "Evaluation of clinical instructors by third year medical students." Academic Medicine **64**(3): 159-164.

Donner-Banzhoff, N., et al. (2003). "Feedback for general practice trainers: developing and testing a standardised instrument using the importance-quality-score method." Medical Education **37**(9): 772-777.

Downing, S. (2003). "Validity: on the meaningful interpretation of assessment data." Medical Education **37**: 830-837.

Downing, S. (2004). "Reliability: on the reproducibility of assessment data." Medical Education **38**: 1006-1012.

Downing, S. M., et al. (1983). "Resident Ratings of Surgical Faculty Improved Teaching Effectiveness Through Feedback." The American Surgeon **49**: 329-332.

Drennan, J. (2003). "Cognitive interviewing: verbal data in the design and pretesting questionnaires." Journal of Advanced Nursing **42**(1): 57-63.

Duffy, B. (2003). "Response order effects - how do people read?" International Journal of Market Research **45**: 457-475.

Elwyn, G. and A. O'Connor (2006). "Develooping a quality criteria framework for patient decision aids: online international Delphi consensus process." British Medical Journal **33**: 417-422.

Field, A. (2009). Discovering statistics using SPSS.

Fink, A. and J. Kosecoff (1984). "Consensus methods: Characteristics and Guidelines for Use." American Journal of Public Health **74**(9): 979-983.

Flin R (2008) Safety at the Sharp end: a guide to non-technical skills Ashgate

Floyd, F. J. and K. F. Widaman (1995). "Factor Analysis in the Development and Refinement of Clinical Assessment Instruments." Psychological Assessment **7**(3): 286 - 299.

Fluit, C., et al. (2010). "Assessing the Quality of Clinical Teachers." Journal of General Internal Medicine **25**(12): 1337-1345.

Fox-Wasylyshyn, S. M. and M. M. El-Masri (2005). "Handling Missing Data in Self-Report Measures." Research in Nursing and Health **28**: 488-495.

GMC (1999). "The Doctor as a Teacher." Retrieved September 2010, from http://www.gmc-uk.org/education/postgraduate/doctor_as_teacher.asp.

GMC (2011). "Developing teachers and trainers in undergraduate medical education." Retrieved May, 2014, from http://www.gmc-uk.org/Developing_teachers_and_trainers_in_undergraduate_medical_education_0511.pdf_56440721.pdf

GMC (2012). "Ready for revalidation. Supporting information for appraisal and revalidation." Retrieved Accessed September 2012, from http://www.gmc-uk.org/static/documents/content/Supporting_information_for_appraisal_and_revalidation.pdf.

Goldstein, A. O. (2005). "Open evaluations are Encouraged." Family Medicine: 386.

GriffithIII, C. H., et al. (1997). "Relationships of How Well Attending Physicians Teach to Their Students' Performances and Residency Choices." Academic Medicine **72**(18): S118-S120.

Guerrasio, J. and M. Weissberg (2012). "Unsigned: why anaonymous evaluations in clinical settings are counterproductive." Medical Education **46**(10): 928-930.

Guyatt, G. H., et al. (1993). "A measurement process for evaluating clinical teachers in internal medicine." CMAJ Canadian Medical Association Journal **149**(8): 1097-1102.

Hargreaves, P. N. and R. Peppiatt (2001). "Is videotaping of consultations acceptable to patients attending a hospice day centre?" Palliative Medicine **15**: 49-54.

Harlen, W. and M. James (1997). "Assessment and Learning: differences and relationships between formative and summative assessment." Assessment in Education **4**(3): 365-379.

Hasson (2000). "Research Guidelines for the Delphi Survey Technique." Journal of Advanced Nursing **32**(4): 1008-1015.

Hauge, L. S., et al. (2001). "The reliability of an instrument for identifying and quantifying surgeons' teaching in the operating room." The American Journal of Surgery **181**: 333-337.

Haycock, A. V., et al. (2010). "Evaluating changes in gastrointestinal endoscopy training over 5 years: closing the audit loop." European Journal of Gastroenterology and Hepatology **22**(3): 368-373.

Hemstokroos, H. H. v. D. and C. P. M. v. d. Vleuten (2005). "Reliability of the Clinical Teaching Effectiveness Instrument " Medical Education **39**: 904-910.

Hsu, C.-C. and B. Sandford (2007). "The Delphi Technique: Making sense of Consensus." Practical assessment,research and evaluation **12**(10): 1-8.

Irby, D. (1983). "Peer Review of Teaching in Medicine." Journal of Medical Education **58**.

Irby, D. and P. Rakestraw (1981). "Evaluating Clinical Teaching in Medicine." Journal of Medical Education **56**: 181-186.

Irby, D. M. (1983). "Evaluating Instruction in Medical Education " Journal of Medical Education **58**: 844-849.

Irby, D. M. (1988). Self assessment inventory for clinical and classroom teaching in medicine. Clinical Teaching of Medical Residents: Roles Teaching and Residents. J. C. Edwards and R. L. Marier. New York, Springer**:** 255-260.

Iwaszkiewicz, M., et al. (2008). "Efforts to Enhance Operating Room Teaching." Journal of Surgical Education **65**(6): 436-440.

Jacobs, R. and S. W. J. Kozlowski (1985). "A Closer Look at Halo Error in Performance Ratings." Academy of Management **28**(1): 201-212.

JAG. "JAG Accreditation System." Retrieved June 2012, 2012, from http://www.jagaccreditation.org/.

JAG (2011). "Joint Advisory Group on Endoscopy." Retrieved September 2010, from http://www.thejag.org.uk/.

JAG (2012). "JAG Endoscopy Training System." Retrieved March, 2012, from http://www.jets.nhs.uk.

Jahangiri, L., et al. (2008). "Assessmnet of Teaching Effectiveness in U.S. Dental Schools and the Value of Triangulation." Journal of Dental Education **72**(6): 707-718.

James, P. A. and J. W. Osborne (1999). "A measure of medical instructional quality in ambulatory settings: the MedIQ." Family Medicine **31**(4): 263-269.

Jamieson, S. (2004). "Likert scales: how to (ab)use them." Medical Education **2004**(38): 1212-1218.

Jobe, J. (2003). "Cognitive psychology and self-reports: Models and methods." Quality of Life Research **12**: 219-227.

Jones, J. and D. Hunter (1995). "Qualitative Research:Consensus methods for medical and health services research." British Medical Journal **311**: 376-380.

Keely, E., et al. (2010). "A teaching encounter card to evaluate clinical supervisors across clerkship rotations." Medical Teacher **32**: e96-e100.

Knafl, K., et al. (2007). "The analysis and interpretation of cognitive interviews for instrument development " Research in Nursing & Health **30**(2): 224-234.

Kruglikova, I., et al. (2010). "The impact of constructive feedback on training in gastrointestinal endoscopy using high-fidelity virtual-reality simulation: a randomised controlled trial." Gut **59**: 181-185.

Lamki, N. and M. Marchand (2006). "The Medical Educator Teaching Portfolio: Its Compilation and Potential Utility." Sultan Qaboos Univ Medical Journal **6**(1): 7-12.

Likert, R. A. (1952). "A technique for the development of attitude scales." Educational and Psychological Assessment **12**: 313-315.

Litzelman, D., et al. (1998). "Factorial validation of a Widely Disseminated Educational Framework for Evaluating Clinical Teachers." Academic Medicine **73**(6): 688-695.

Litzelman, D. K., et al. (1998). "Beneficial and Harmful Effects of Augmented Feedback on Physicians' Clinical- teaching Performances." Academic Medicine **73**(3): 324-332.

Longmore, M., et al. (2001). Oxford Handbook of Clinical Medicine, Oxford University Press.

Macdougall, C. and C. O'Halloran (2001). "Keeping it simple - audio taping in consultation performance assessment." Medical Education **35**: 1091.

Mahler, S. and D. E. Benor (1984). "Short and longterm effects of a teacher-training workshop in medical school." Higher Education **13**(3): 265 -273.

Mahmood, T. and A. Darzi (2004). "The learning curve for a colonoscopy simulator in the absence of any feedback." Surgical Endoscopy **18**: 1224-1230.

Maker, V. K., et al. (2004). "Are you a Surgical Role Model?" Current Surgery **61**(1): 111-115.

Maker, V. K., et al. (2004). "Faculty Evaluations: Diagnostic and Therapeutic." Current Surgery **61**(6): 597 - 598.

Maker, V. K., et al. (2006). "Ongoing Faculty Evaluations: Developmental Gain or just more Pain?" Current Surgery **63**(1): 80 - 84.

Maslow, A. H. (1970). Motivation and Personality. New York, Harper Collins.

Mathers, N. J., et al. (1999). "Portfolios in continuing Medical Education - Effective and Efficient?" Medical Education **33**(7): 521-530.

McGrath, C., et al. (2005). "Development and evaluation of a questionnaire to evaluate clinical dental teachers (ECDT)." British Dental Journal **198**(1): 45-48.

Mckee, M., et al. (1191). "How Representative are Members of Expert Panels?" Quality Assurance in Health Care **3**(2): 84-94.

McLean, M. (2001). "Rewarding reaching excellence. Can we measure teaching 'excellence'? Who should be the judge?" Medical Teacher **23**(1): 6-11.

McLean, M. (2001). "Rewarding teaching excellence. Can we measure teaching 'excellence'? Who should be the judge?" Medical Teacher **23**(1): 6-11.

McLean, M. (2001). "Rewarding Teaching Excellence. Can we measure teaching 'excellence'? Who should be the judge?" Medical Teacher **23**(1): 6-11.

Mehta, T., et al. (2011). "Development and roll out of the JETS e-portfolio: a web based electronic portfolio for endoscopists." Frontline Gastroenterology **2**: 35-42.

Menachery, E. P., et al. (2006). "Physician Characteristics Associated with Proficiency in Feedback Skills." Journal of General Internal Medicine **21**: 440-446.

Merriam, S. B. (2007). Chapter 11: Traditional Learning Theories. Learning in adulthoof: a comprehensive guide. J. Wiley, Jossey-Bass.

Microsoft (2007). Excel 2007. Redmond, Microsoft. **12**.

Miller, K. (2003). "Conducting Cognitive Interviews to Understand Question-Response Limitations." American Journal of Health Behavior **27**(3): 264-272.

Molodysky, E., et al. (2006). "Identifying and Training Effective Clinical Teachers." Australian Family Physician **35**(1): 53-55.

Moore, A. (2000). Teaching and Learning: Pedagogy, curriculum and culture. London, Routledge Falmer.

Morrison, E. H., et al. (2002). "Reliability and Validity of an Objective Structured Teaching Examination for Generalist Resident Teachers." Academic Medicine **77**(10): S29 - S32.

Morrison, J. (2003). "ABC of learning and teaching in medicine. Evaluation." British Medical Journal **326**: 385-387.

Murry, J. and J. O. Hammons (1995). "Delphi: A Versatile Methodology for conducting Qualitative Research." The Review of Higher Education **18**(4): 423-436.

Murtagh, F. (2007). "The value of cognitive interviewing techniques in palliative care research." Palliative Medicine **21**: 87-93.

Nishisato, N. and Y. Torii (1970). "Effects of categorising continous normal distributions on the product-moment correlation." Japanese Psychological Research **13**: 45-49.

Okoli, C. and S. Pawlowski (2004). "The Delphi method as a research tool: an example, design coniderations and applications." Information and management **42**: 15-29.

Perlberg, A., et al. (1972). "Microteaching and Videotape Recordings: A new Approach to Improving Teaching." Journal of Medical Education **47**: 43-50.

Peyton, J. (1998). The learning cycle. Teaching and learning in medical practice. J. Peyton. Manticore, Silver Birches**:** 13 - 19.

Pill, J. (1971). "The Delphi Method: substance, context, a critique and an annotated bibliography." Socioeconomic planning sciences **5**(1): 57-71.

Ramani, S. and S. Leinster (2008). "AMEE Guide no. 34: Teacing in the clinical environment." Medical Teacher **30**: 347-364.

Rees, C. and M. Rutter (2010). "NREG: old banger or new vehicle for research?" Frontline Gastroenterology **1**: 59-62.

Richardson, B. K. (2004). "Feedback." Academic Emergency Medicine **11**(12): 1283.e1281-e1285.

Risucci, D. A., et al. (1992). "Reliability and accuracy of resident evaluations of surgical faculty." Evaluation and the Health Professions **15**(3): 313-324.

Roff, S., et al. (2005). "Development and validation of an instrument to measure the postgraduate clinical learning and teaching environment for hospital-based junior doctors in the UK." Medical Teacher **27**(4): 326-331.

Rolfe, I. and J. McPherson (1995). "Formative assessment:how am I doing?" The Lancet **345**: 837-839.

Rosenshine, B. (1975). Teaching Effectiveness: Its Meaning, Assessment and Improvement. R. E. Hull. New Jersey, Educational Technology Publications**:** 105-120.

Rust, J. and S. Golombok (2009). Modern Psychometrics The science of Pyshcometric assessment, Routledge.

Sarker, S. K., et al. (2005). "Assessing the teaching of technical skills." The American Journal of Surgery **189**: 416-418.

Schultz, K. and D. Latif (2006). "The Planning and Implementation of a Faculty Peer Review Teaching Project." American Journal of Pharmaceutical Education **70**(2): 1-6.

Schum, T. R. and K. J. Yindra (1996). "Relationship between Systematic Feedback to Faculty and Ratings of Clinical Teaching." Academic Medicine **71**(10): 1100-1102.

Schuwirth, L. W. T. and C. P. M. V. d. Vleuten (2006). "A plea for new psychometric models in educational assessment." Medical Education **40**: 296 - 300.

Schwarz, N. and D. Oyserman (2001). "Asking Questions About Behavior; Cognition, Communication, and Questionnaire Construction." American Journal of Evaluation **22**(2): 127-160.

Sepucha, K., et al. (2007). "An approach to measuring the quality of breast cancer decisions." Patient Education & Counseling **65**(2): 261-269.

Shavelson, R. J. and N. M. Webb (1991). Generalizability Theory A Primer. Thousand Oaks California, Sage Publications.

Siddiqui, Z. S., et al. (2007). "Twelve tips for the observation of teaching." Medical Teacher **29**: 297-300.

Sim, D. J., et al. (2004). "Effects of the European Working Time Directive on anaesthetic training in the United Kingdom." Anaesthesia **59**(8): 781-784.

Sinha, I. P., et al. (2011). "Using the Delphi Technique to Determine Which Outcomes to Measure in Clinical Trials: Recommendations for the Future Based on a Systematatic Review of Existing Studies." PLoS Medicine **8**(1): 1-5.

Sircus, W. (2003). "Milestones in the evolution of endoscopy: a short history." Journal of the Royal College of Physicians Edinburgh **33**: 124-134.

Skeff, K. M., et al. (1986). "Evaluation of the seminar method to improve clinical teaching." Journal of General Internal Medicine **1**(5): 315-322.

Snell, L., et al. (2000). "A review of the evaluation of clinical teaching: new perspectives and challenges." Medical Education **34**(10): 862-870.

Speer, A. J. and D. M. Elnicki (1999). "Assessing the Quality of Teaching." The American Journal of Medicine **106**: 381-384.

Spencer, J. (2003). "Learning and teaching in the clinical environment." British Medical Journal **326**: 591-594.

SPSS (2005). SPSS for Windows Version 14. Chicago.

SPSS (2005). SPSS for Windows Version 14. Chicago.

Stalmeijer, R., et al. (2010). "Combined student ratings and self assessment provide useful feedback for clinical teachers." Advances in Health Science Education **15**: 315-328.

Stalmeijer, R., et al. (2009). "Cognitive apprenticeship in clinical practice: can it stimulate learning in the opinion of students." Advances in Health Science Education **14**: 535-546.

Stalmeijer, R. E., et al. (2008). "The development of an instrument for evaluating clinincal teachers: involving stakeholders to deterrmine content validity." Medical Teacher **30**: 272-277.

StatsToDo (2013). "Stats To Do." Retrieved April 2013, from https://www.statstodo.com/SSizCorr_Pgm.php.

Steiner, I. P., et al. (2000). "Faculty Evaluation by Residents in an Emergency Medicine Program: A New Evaluation Instrument." Academic Emergency Medicine **7**(9): 1015-1021.

Steinert, Y., et al. (2006). "A systematic review of faculty development initiatives designed to improve teaching effectiveness in medical education: BEME Guide No. 8." Medical Teacher **28**(6): 497-526.

Stern, D. T., et al. (2000). "Is There a Relationship betwene Attending Physicians' and Residents' Teaching Skills and Students' Examination Scores." Academic Medicine **75**(11): 1144-1146.

Stewart, J. and C. O'Halloran (1999). "Identifying appropriate tasks for the preregistration year: modified Delphi technique." British Medical Journal **319**: 224-229.

Stone, S., et al. (2002). "Development and Implementation of an Objective Structured Teaching Exercise (OSTE) to evaluate Improvement in Feedback skills Following a Faculty Development Workshop." Teaching and Learning in Medicine **15**(1): 7 - 13.

Streiner, D. and G. R. Norman (2008). Health Measurement Scales, Oxford University Press.

Sullivan, P. B., et al. (2012). "Peer observation of teaching as va faculty development tool." BMC Medical Education **12**(26).

Surveymonkey (2008). "Survey monkey." Retrieved September 2010, from www.surveymonkey.com.

Sutkin, G., et al. (2008). "What Makes a Good Clinical Teacher in Medicine? A review of the Literature." Academic Medicine **83**(5): 452-457.

Teague, R., et al. (2002). "Setting standards for colonoscopic teaching and training " Journal of Gastroenterology and Hepatology **17** (Suppl.): S50 - S53

Thomas-Gibson, S., et al. (2007). "Intensive training over 5 days improves colonoscopy skills long-term." Endoscopy **39**: 818-824.

Thuraisingam, A., et al. (2006). "Insights into endoscopy training: a qualitative study of learning experience." Medical Teacher **28**(5): 453-459.

Tiberius, R. G., et al. (1989). "The influence of Student Evaluative Feedback on the improvement of Clinical Teaching " The Journal of Higher Education **60**(6): 665-681.

Tortolani, A. J., et al. (1991). "Resident Evaluation of Surgical Faculty." Journal of Surgical Research **51**: 186-191.

Trowbridge, R. L., et al. (2011). "A systematic review of the use and effectiveness of the Objective Structured Teaching Encounter." Medical Teacher **33**: 893 - 903.

Villiers, M. R. d., et al. (2005). "The Delphi Technique in health Sciences education research." Medical Teacher **27**(7): 639-643.

Vygotsky, L. S. (1978). Mind in society: The development of higher psychological processes. Cambridge, Massachetts, Harvard University Press.

Wamsley, M. A., et al. (2004). "A Literature Review of 'Resident-as-Teacher' Curricula. Do Teaching Courses Make a Difference?" Journal of General Internal Medicine **19**: 574-581.

Wells, C. (2008). High quality teaching of endoscopy: Defining the attributes of the trainer. School of Medical Education, Newcastle-upon-Tyne University. **MD**.

Wells, C. W., et al. (2009). "Trainees in gastroenterology views on teaching in clinical gastroenterology and endoscopy." Medical Teacher **31**: 138-144.

Wildly, H. and S. Clarke (2009). "Using Cognitive interviews to pilot an international survey of principal preparation: A Western Australian perspective." Educ Asse Evall Acc **21**: 105-117.

Windish, D. M., et al. (2004). "Clinician-teachers' Self-assessments Versus Learners' Perceptions." Journal of General Internal Medicine **19**: 554-557.

Wood, D., et al. (1976). "The role of tutoring in problem solving." Journal of child psychology and psychiatry **17**: 89-100.

Yammarino, F. J., et al. (1991). "Understanding Mail Survey Response Behaviour." Public Opinion Quarterly **55**: 613-639.

Yeates, P., et al. (2008). "What can we expect of clinical teachers? Establishing consensus on applicable skills, attitudes and practices." Medical Education **42**: 134-142.

Attributes as described by Wells (2010)

A Preparation
1. Everyone's roles with respect to training were clear
2. The nursing staff were not inconvenienced by the list
3. The trainer agreed rules for the teaching with the trainee
4. The list was populated with cases appropriate to the needs of the trainee (in terms of volume and nature of cases)
5. The trainer agreed SMART goals for the session with his trainee at the start of the list

B The learning atmosphere
6. The trainer made the trainee feel welcome
7. The trainer ensured that the trainee was physically comfortable
8. The trainer was approachable and it was easy to ask him questions
9. The trainer was honest with the trainee
10. The trainer showed respect for the trainee
11. The trainer built the trainee's self-belief and confidence
12. The trainer was patient and calm
13. The trainer had realistic expectations of his trainee

C Modelling
14. The trainer always ensured that the patient was comfortable and safe
15. The trainer dealt with any complications (if any)
16. The trainer was able to describe how he performed any endoscopic manoeuvres to his trainee that was understandable to the trainee
17. The trainer demonstrated how to do a procedure to the trainee where necessary

D Coaching
18. The trainer taught the trainee to handle the scope gently
19. The trainer used other equipment that can support teaching (e.g. the magnetic imager, models etc) appropriately
20. The trainer concentrated on one thing at a time and did not overburden the trainee

21. The trainer intervened in a timely fashion if the trainee was failing to make progress (either at a predefined time or if the trainee is struggling excessively)
22. The trainer dealt with any mistakes made by the trainee
23. The trainer used a mix of suggestions, prompts, solutions & instructions to the trainee where appropriate
24. The trainer closely observed the process and was aware of what the trainee was doing
25. The trainer repositioned the trainee when appropriate

E Scaffolding
26. The trainer's interventions were proportional to the competence of the trainee and the difficulty of the procedure
27. The trainer was appropriately available for the trainee
28. The style and pace of the training was appropriate to the trainee
29. The trainer adjusted the position he was standing in the room appropriately, withdrawing as the trainee progressed

F Articulation
30. The trainer agreed a common vocabulary with the trainee
31. The trainer helped the trainee to assimilate all the information (from feel, screen, patient comfort, nurses) to help them progress
32. The quantity of dialogue was appropriate for the trainee and the specific teaching episode
33. The trainer provided opportunities for the trainee to speak and actively listened to the trainee
34. The trainer reassured the trainee at appropriate times
35. The trainer questioned the trainee at appropriate times
36. The trainer provided explanations at appropriate times
37. The trainer checks that the trainee has understood what has been said
38. The trainer asked the trainee to verbally run through a manoeuvre before doing it when appropriate
39. The trainer used non verbal communication positively
40. The trainer asked the trainee to demonstrate the problem when appropriate

G Exploration
41. The trainer allowed the trainee enough time to carry out the procedure without rushing them
42. The trainer helped the trainee to carry out the procedure independently whilst ensuring patient safety
43. The trainer allowed the trainee to learn by trial and error provided patient safety was not compromised
44. The trainer allowed the trainee to find their limits and to understand when they should give up trying
45. The trainer handed back the scope after overcoming a difficult manoeuvre

H Reflection and Feedback

46. The trainer provided feedback close to the teaching event
47. The trainer delivered the feedback in a framework appropriate for the trainee
48. The trainer helped the trainee reflect on their performance
49. The trainer reinforced positive aspects of the trainee's performance
50. The trainer identified aspects for the trainee to develop and improve
51. The trainer dealt with any lack of insight in the trainee
52. The trainer helped the trainee to asses if the goals for the session had been achieved
53. The trainer and trainee agreed goals for future sessions
54. The trainer challenged the trainee to justify what they had done or were about to do when appropriate

I Global

55. The trainer scheduled enough training lists for the trainee
56. The trainer limited the number of trainees he was teaching to a level that each trainee received adequate training
57. The trainer agreed the rules of the training and was consistent in the application of these rules
58. The trainer gets to know the trainee personally e.g. career aims, family etc
59. The trainer developed a good working relationship with the trainee
60. The trainer and trainee agreed and worked towards common goals during the training period with a long term training plan
61. The trainer role modelled the desired behaviours of an endoscopist
62. The trainer is credible as an endoscopist and respected by trainee
63. The trainer is proficient & experienced in the endoscopy procedure he is teaching
64. The trainer collected data to use as feedback to the trainee eg DOPS forms, CuSum etc

J Content

65. The trainer taught according to the guidelines as per JAG and the DOPS
66. The trainer understands the mechanics of endoscopy
The trainer taught the trainee to handle the scope gently appears to disappear and reappear!
67. The trainer demonstrated their competence at performing endoscopy
68. The trainer taught the importance of feel (tactile feedback)
69. The trainer had a broad knowledge about the practice of endoscopy
70. The trainer taught the trainee to thoroughly examine the mucosa
71. The trainer taught the trainee to keep the lumen in view
72. The trainer taught the trainee to keep insufflation to a minimum
73. The trainer taught the trainee about loop resolution
74. The trainer taught the whole process of endoscopy to the trainee
75. The trainer taught the importance of fine endoscopic control
76. The trainer taught the trainee to communicate with the nurses

77. The trainer taught the trainee to communicate with the patient

K Heuristic strategies

78. The trainer took advantage of opportune moments to teach
79. The trainer taught specific non standard strategies from his own experience when he felt this would help the trainee
80. The trainer taught the theory of endoscopy first
81. The trainer made the trainee aware of how the equipment worked
82. The trainer taught the basics of endoscopy (consent, sedation, how the scope moves)
83. The trainer let the trainee handle the endoscope outside of the patient before using the scope on a patient
84. The trainer gradually increased the difficulty of the tasks set for the trainee
85. The trainer adhered to the learning plan and reviewed the long term progress of the trainee
86. The trainer knows the learning goals of the trainee and works towards these goals
87. The trainer provided continued supervision for his former trainee even when the trainee was fully trained

Appendix 2 List of the attributes following the cognitive interviews

A Preparation

1. The trainer clarifies everyone's role before a training encounter so that each individual knows how they are involved in the training process.
2. The nurses are informed it is a training list to ensure they are supportive of the trainee.
3. The trainer prepares the endoscopy training lists to meet the current needs of his trainee, both in volume and the nature of the cases on the list
4. The trainer agreed SMART goals for the session with his trainee at the start of the list. (S = specific, M = measurable, A = achievable, R = relevant, T = can be achieved in the timeframe)
5. The trainer planned enough time for feedback
6. The trainer agreed the rules of the training and was consistent in the application of these rules

B The learning atmosphere

7. The trainer made the trainee feel welcome
8. The trainer ensured that the trainee was physically comfortable (including neither being tired or in actual physical discomfort).
9. The trainer was approachable and it was easy to ask him questions
10. The trainer acknowledged when he was unable to explain the manoeuvres he had performed. – move to articulation
11. The trainer showed respect for the trainee
12. The trainer built the trainee's self-belief and confidence
13. The trainer was patient and calm
14. The trainer had realistic expectations of his trainee

C Modelling

14. The trainer always ensured that the patient was comfortable and safe and their dignity was maintained.
15. The trainer dealt with any complications (if any)
16. The trainer was able to describe how he performed any endoscopic manoeuvres to his trainee that was understandable to the trainee and the trainee is left with an appreciation of how to perform the procedure.
17. The trainer demonstrated how to do a procedure to the trainee where necessary

D Coaching

18. The trainer taught the trainee to handle the scope gently
19. The trainer used teaching aids that can support learning (e.g. the magnetic imager, diagrams, models etc)
20. The trainer concentrated on one thing at a time and did not overburden the trainee

21. The trainer intervened in a timely fashion if the trainee was failing to make progress (either at a predefined time or if the trainee is struggling excessively)
22. The trainer dealt with any slips, errors or mistakes made by the trainee
23. The trainer used a mix of suggestions, prompts, solutions & instructions to the trainee where appropriate
24. The trainer closely observed the process and was aware of what the trainee was doing
25. The trainer physically moved his trainee to help them to achieve the desired outcome.

E Scaffolding
26. The trainer's interventions were proportional to the competence of the trainee and the difficulty of the procedure
27. The trainer was appropriately available for the trainee
28. The style and pace of the training was appropriate to the trainee
29. The trainer adjusted the position he was standing in the room appropriately, withdrawing as the trainee progressed

F Articulation
30. The trainer agreed a common vocabulary with the trainee
31. The trainer helped the trainee to assimilate all the information (from feel, screen, patient comfort, nurses) to help them progress
32. The quantity of dialogue was appropriate for the trainee and the specific teaching episode
33. The trainer provided opportunities for the trainee to speak and actively listened to the trainee
34. The trainer reassured the trainee at appropriate times
35. The trainer questioned the trainee at appropriate times
36. The trainer provided explanations at appropriate times
37. The trainer checks that the trainee has understood what has been said
38. The trainer asked the trainee to verbally run through a manoeuvre before doing it when appropriate
39. The trainer is aware of how non-verbal signals may affect the trainee
40. The trainer asked the trainee to demonstrate the problem when appropriate

G Exploration
41. The trainer allowed the trainee enough time to carry out the procedure without rushing them
42. The trainer helped the trainee to carry out the procedure independently whilst ensuring patient safety
43. The trainer allowed the trainee to attempt procedures independently learning from what does and does not work as long as patient safety is not compromised.
44. The trainer allowed the trainee to find their limits and to understand when they should give up trying

45. The trainer handed back the scope after overcoming a difficult manoeuvre
46. The trainer and trainee agreed goals for future sessions (moved from 53)

H Reflection and Feedback

47. The trainer provided feedback close to the teaching event
48. The trainer delivered the feedback in a structure appropriate to the trainee
49. The trainer helped the trainee reflect on their performance
50. The trainer reinforced positive aspects of the trainee's performance
51. The trainer identified aspects for the trainee to develop and improve
52. The trainer helped the trainee to asses if the goals for the session had been achieved
53. The trainer used appropriate challenges (e.g. to justify a manoeuvre they have performed) to help the trainee progress

I Global

54. The trainer scheduled enough training lists for the trainee
55. The trainer limited the number of trainees he was teaching to ensure that each trainee received adequate training
56. The trainer gets to know the trainee personally e.g. career aims, family etc
57. The trainer developed a good working relationship with the trainee
58. The trainer and trainee agreed and worked towards common goals during the training period with a long term training plan
59. The trainer set a good professional example through their own behaviour
60. The trainer collected data to use as feedback to the trainee eg DOPS forms, CuSum etc

J Content

65. The trainer taught according to the guidelines as per JAG and the DOPS
66. The trainer can explain the mechanics of endoscopy
67. The trainer demonstrated their competence at performing endoscopy
68. The trainer taught the importance of feel (tactile feedback)
69. The trainer had a broad knowledge about the practice of endoscopy
70. The trainer taught the trainee to thoroughly examine the mucosa
71. The trainer taught the trainee to keep the lumen in view
72. The trainer taught the trainee to keep insufflation to a minimum
73. The trainer taught the whole process of endoscopy to the trainee e.g. the indications, consent and communication
74. The trainer emphasised the importance of fine tip control.
75. The trainer encouraged the trainee to communicate appropriately with the nurses
76. The trainer encouraged the trainee to communicate appropriately with the patient

K Heuristic strategies

78. The trainer took advantage of opportune moments to teach

79. The trainer taught specific non standard strategies from his own experience when he felt this would help the trainee
80. The trainer taught the theory of endoscopy before each new stage
81. The trainer made the trainee aware of how the equipment worked
82. The trainer gradually increased the difficulty of the tasks set for the trainee
83. The trainer adhered to the learning plan and reviewed the long term progress of the trainee
84. The trainer knows the learning goals of the trainee and works towards these goals

Appendix 3 Questionnaire for Round 1 of the Delphi

Delphi study: attributes of a 'high-quality' trainer

1. Attributes of an endoscopy trainer: Instructions

Thank you for agreeing to take part in this research study.

We are aiming to create an 'evaluation toolkit' which can be used to give formative feedback to a trainer regarding their teaching of endoscopy. This scientifically derived toolkit aims to replace the current DOTS tool developed through expert consensus.

This questionnaire contains 87 questions and will take a maximum of 45 minutes to complete however as long as you enter it through the same computer you do not need to complete it all in one go. This is the first of 3 rounds of questionnaires. We appreciate that this may be difficult to complete in your busy working lives but the aim is that this will then result in a workable tool that will improve endoscopic training in the UK.

The evaluation toolkit will have two components:

- DOTS (a direct observation of teaching skills) used to evaluate the single teaching encounter
- LETS (long-term evaluation of teaching skills) which can assess attributes that are displayed over the longer term.

For instance a DOTS would be completed after a single list whereas a LETS would be completed at the end of a trainee's hospital attachment.

Although we do not wish to specify the number of statements in the final tool please remember we want this tool to be as user-friendly as possible, i.e. usable after each endoscopy list (DOTS), and at the end of an attachment (LETS).

INSTRUCTIONS

ROUND 1

• This questionnaire contains a list of attributes describing the 'high-quality' endoscopy trainer derived from previous research
• For each attribute please rate its suitability for the DOTS and then the LETS on a five-point scale (strongly disagree to strongly agree); the attribute may be suitable to one, both or neither.
• If you feel the attribute's suitability could be improved please suggest any modifications.
• Please also feel free to make comments which explain the reasoning for your score
• If after rating all the items you feel we have missed any key aspects of a 'high-quality' trainer please feel free to suggest new statements.
• When rating the statements please apply the following criteria; each attribute should be
o Measurable
o Assessable by peer and trainee
o Generic to all procedures
o Generic to all levels of trainee
o And, for the DOTS, likely to occur in every training list

All completed questionnaires will be anonymised and no identifiable data will be included in the summary of opinions or any publications regarding this study.

If we haven't heard from you we will send you two reminder emails. Participation in this study is entirely voluntary; if you wish to withdraw from the study at any point we request you return a blank questionnaire or email the address below and we will not trouble you again.

If you have any further questions about the study or require assistance completing the questionnaire please contact the research team at louise.macdougall@nhct.nhs.uk

Many thanks once again for completing this questionnaire

Delphi study: attributes of a 'high-quality' trainer

Dr Louise Macdougall

On behalf of
Professor Roger Barton
Dr Sally Corbett
Dr Mark Welfare
Dr Chris Wells

N.B. This project has been discussed with and is supported by the Endoscopy Training Leads across the UK

Delphi study: attributes of a 'high-quality' trainer

2. Attributes of an endoscopy trainer: Demographics

We would be grateful to know a little more about you.

1. Name

2. Sex:

- ◯ Male
- ◯ Female

3. Profession:

- ◯ Surgeon
- ◯ Physician
- ◯ Nurse

4. Have you attended a 'Training the trainers' course?

- ◯ Yes
- ◯ No

5. For trainees only: What year of training are you in?

- ◯ Year 1
- ◯ Year 2
- ◯ Year 3
- ◯ Year 4
- ◯ Year 5

Delphi study: attributes of a 'high-quality' trainer

3. Attributes of an endoscopy trainer: preparation

6. The trainer clarifies everyone's role before a training encounter so that each individual knows how they are involved in the training process.

This item should be included in the...

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
DOTS	◯	◯	◯	◯	◯
LETS	◯	◯	◯	◯	◯

Please suggest any modifications/justify your answer (optional)

7. The nurses are informed it is a training list to ensure they are supportive of the trainee.

This item should be included in the...

	Strongly disagree	Diagree	Neutral	Agree	Strongly agree
DOTS	◯	◯	◯	◯	◯
LETS	◯	◯	◯	◯	◯

Please suggest any modifications/justify your answer (optional)

8. The trainer prepares the endoscopy training lists to meet the current needs of his trainee, both in volume and the nature of the cases on the list.

This item should be included in the...

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
DOTS	◯	◯	◯	◯	◯
LETS	◯	◯	◯	◯	◯

Please suggest any modifications/ justify your answer (optional)

Delphi study: attributes of a 'high-quality' trainer

9. The trainer agreed SMART goals for the session with his trainee at the start of the list. (S = specific, M = measurable, A = achievable, R = relevant, T = can be achieved in the timeframe).

This item should be included in the...

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
DOTS	◯	◯	◯	◯	◯
LETS	◯	◯	◯	◯	◯

Please suggest any modifications/ justify your answer (optional)

10. The trainer agreed the rules of the training and was consistent in the application of these rules.

This item should be included in the....

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
DOTS	◯	◯	◯	◯	◯
LETS	◯	◯	◯	◯	◯

Please suggest any modifications/ justify your answer (optional)

11. The trainer planned enough time for feedback.

This item should be included in the...

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
DOTS	◯	◯	◯	◯	◯
LETS	◯	◯	◯	◯	◯

Please suggest any modifications/ justify your answer (optional)

Delphi study: attributes of a 'high-quality' trainer

4. Attributes of an endoscopy trainer: the learning atmosphere

12. The trainer made the trainee feel welcome.

This item should be included in the...

	Strongly disagree	Disagree	Neutral	Agree	Strongly Agree
DOTS	◯	◯	◯	◯	◯
LETS	◯	◯	◯	◯	◯

Please suggest any modifications/justify your answer (Optional)

13. The trainer ensured that the trainee was physically comfortable (including neither being tired or in actual physical discomfort).

This item would be suitable for....

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
DOTS	◯	◯	◯	◯	◯
LETS	◯	◯	◯	◯	◯

Please suggest any modifications/justify your answer (optional)

14. The trainer was approachable and it was easy to ask him questions.

This item would be suitable for....

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
DOTS	◯	◯	◯	◯	◯
LETS	◯	◯	◯	◯	◯

Please suggest any modifications/justify your answer (optional)

Delphi study: attributes of a 'high-quality' trainer

15. The trainer showed respect for the trainee.

This item would be suitable for....

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
DOTS	◯	◯	◯	◯	◯
LETS	◯	◯	◯	◯	◯

Please suggest any modifications/justify your answer (optional)

16. The trainer built the trainee's self-belief and confidence.

This item would be suitable for....

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
DOTS	◯	◯	◯	◯	◯
LETS	◯	◯	◯	◯	◯

Please suggest any modifications/justify your answer (optional)

17. The trainer was patient and calm.

This item would be suitable for....

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
DOTS	◯	◯	◯	◯	◯
LETS	◯	◯	◯	◯	◯

Please suggest any modifications/justify your answer (optional)

18. The trainer had realistic expectations of his trainee.

This item would be suitable for....

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
DOTS	◯	◯	◯	◯	◯
LETS	◯	◯	◯	◯	◯

Please suggest any modifications/justify your answer (optional)

Delphi study: attributes of a 'high-quality' trainer

5. Attributes of an endoscopy trainer: Modelling

19. The trainer always ensured that the patient was comfortable and safe and their dignity was maintained.

This item should be included in the...

	Strongly disagree	Disagree	Neutral	Agree	Strongly Agree
DOTS	◯	◯	◯	◯	◯
LETS	◯	◯	◯	◯	◯

Please suggest any modifications/justify your answer (Optional)

20. The trainer dealt with any complications (if any).

This item should be included in the...

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
DOTS	◯	◯	◯	◯	◯
LETS	◯	◯	◯	◯	◯

Please suggest any modifications/justify your answer (Optional)

21. The trainer was able to describe how he performed any endoscopic manoeuvres to his trainee that was understandable to the trainee and the trainee is left with an appreciation of how to perform the procedure.

This item should be included in the...

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
DOTS	◯	◯	◯	◯	◯
LETS	◯	◯	◯	◯	◯

Please suggest any modifications/justify your answer (Optional)

Delphi study: attributes of a 'high-quality' trainer

22. The trainer demonstrated a procedure to the trainee where necessary.

This item should be included in the...

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
DOTS	◯	◯	◯	◯	◯
LETS	◯	◯	◯	◯	◯

Please suggest any modifications/justify your answer (Optional)

Page 9

Delphi study: attributes of a 'high-quality' trainer

6. Attributes of an endoscopy trainer: Coaching

23. The trainer taught the trainee to handle the scope gently.

This item should be included in the...

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
DOTS	◯	◯	◯	◯	◯
LETS	◯	◯	◯	◯	◯

Please suggest any modifications/justify your answer (Optional)

24. The trainer used teaching aids that can support learning (e.g. the magnetic imager, diagrams, models etc).

This item should be included in the...

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
DOTS	◯	◯	◯	◯	◯
LETS	◯	◯	◯	◯	◯

Please suggest any modifications/justify your answer (Optional)

25. The trainer concentrated on one thing at a time and did not overburden the trainee.

This item should be included in the...

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
DOTS	◯	◯	◯	◯	◯
LETS	◯	◯	◯	◯	◯

Please suggest any modifications/justify your answer (Optional)

Delphi study: attributes of a 'high-quality' trainer

26. The trainer intervened in a timely fashion if the trainee was failing to make progress (either at a predefined time or if the trainee is struggling excessively).

This item should be included in the...

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
DOTS	◯	◯	◯	◯	◯
LETS	◯	◯	◯	◯	◯

Please suggest any modifications/justify your answer (Optional)

27. The trainer dealt with any slips, errors or mistakes made by the trainee.

This item should be included in the...

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
DOTS	◯	◯	◯	◯	◯
LETS	◯	◯	◯	◯	◯

Please suggest any modifications/justify your answer (Optional)

28. The trainer used a mixture of suggestions, prompts, solutions and instructions to the trainee where appropriate.

This item should be included in the...

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
DOTS	◯	◯	◯	◯	◯
LETS	◯	◯	◯	◯	◯

Please suggest any modifications/justify your answer (Optional)

Delphi study: attributes of a 'high-quality' trainer

29. The trainer closely observed the process and was aware of what the trainee was doing.

This item should be included in the...

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
DOTS	◯	◯	◯	◯	◯
LETS	◯	◯	◯	◯	◯

Please suggest any modifications/justify your answer (Optional)

30. The trainer physically moved his trainee to help them to achieve the desired outcome.

This item should be included in the...

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
DOTS	◯	◯	◯	◯	◯
LETS	◯	◯	◯	◯	◯

Please suggest any modifications/justify your answer (Optional)

Delphi study: attributes of a 'high-quality' trainer

7. Attributes of an endoscopy trainer: Scaffolding

31. The trainer's interventions were proportional to the competence of the trainee and the difficulty of the procedure.

This item should be included in the...

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
DOTS	◯	◯	◯	◯	◯
LETS	◯	◯	◯	◯	◯

Please suggest any modifications/justify your answer (Optional)

32. The trainer was appropriately available for the trainee.

This item should be included in the...

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
DOTS	◯	◯	◯	◯	◯
LETS	◯	◯	◯	◯	◯

Please suggest any modifications/justify your answer (Optional)

33. The style and pace of the training was appropriate to the trainee.

This item should be included in the...

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
DOTS	◯	◯	◯	◯	◯
LETS	◯	◯	◯	◯	◯

Please suggest any modifications/justify your answer (Optional)

Delphi study: attributes of a 'high-quality' trainer

34. The trainer adjusted the position he was standing in the room appropriately, withdrawing as the trainee progressed.

This item should be included in the...

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
DOTS	◯	◯	◯	◯	◯
LETS	◯	◯	◯	◯	◯

Please suggest any modifications/justify your answer (Optional)

Delphi study: attributes of a 'high-quality' trainer

8. Attributes of an endoscopy trainer: Articulation

35. The trainer agreed a common vocabulary with the trainee.

This item should be included in the...

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
DOTS	◯	◯	◯	◯	◯
LETS	◯	◯	◯	◯	◯

Please suggest any modifications/justify your answer (Optional)

36. The trainer helped the trainee to assimilate all the information (from feel, screen, patient comfort, nurses) to help them progress.

This item should be included in the...

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
DOTS	◯	◯	◯	◯	◯
LETS	◯	◯	◯	◯	◯

Please suggest any modifications/justify your answer (Optional)

37. The quantity of dialogue was appropriate for the trainee and the specific teaching episode.

This item should be included in the...

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
DOTS	◯	◯	◯	◯	◯
LETS	◯	◯	◯	◯	◯

Please suggest any modifications/justify your answer (Optional)

Delphi study: attributes of a 'high-quality' trainer

38. The trainer provided opportunities for the trainee to speak and actively listened to the trainee.

This item should be included in the...

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
DOTS	◯	◯	◯	◯	◯
LETS	◯	◯	◯	◯	◯

Please suggest any modifications/justify your answer (Optional)

39. The trainer reassured the trainee at appropriate times.

This item should be included in the...

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
DOTS	◯	◯	◯	◯	◯
LETS	◯	◯	◯	◯	◯

Please suggest any modifications/justify your answer (Optional)

40. The trainer questioned the trainee at appropriate times.

This item should be included in the...

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
DOTS	◯	◯	◯	◯	◯
LETS	◯	◯	◯	◯	◯

Please suggest any modifications/justify your answer (Optional)

41. The trainer provided explanations at appropriate times.

This item should be included in the...

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
DOTS	◯	◯	◯	◯	◯
LETS	◯	◯	◯	◯	◯

Please suggest any modifications/justify your answer (Optional)

Delphi study: attributes of a 'high-quality' trainer

42. The trainer checks that trainee has understood what has been said.

This item should be included in the...

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
DOTS	◯	◯	◯	◯	◯
LETS	◯	◯	◯	◯	◯

Please suggest any modifications/justify your answer (Optional)

43. The trainer asked the trainee to verbally run through a manoeuvre before doing it when appropriate.

This item should be included in the...

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
DOTS	◯	◯	◯	◯	◯
LETS	◯	◯	◯	◯	◯

Please suggest any modifications/justify your answer (Optional)

44. The trainer is aware of how non-verbal signals may affect the trainee.

This item should be included in the...

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
DOTS	◯	◯	◯	◯	◯
LETS	◯	◯	◯	◯	◯

Please suggest any modifications/justify your answer (Optional)

45. The trainer asked the trainee to demonstrate the problem when appropriate.

This item should be included in the...

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
DOTS	◯	◯	◯	◯	◯
LETS	◯	◯	◯	◯	◯

Please suggest any modifications/justify your answer (Optional)

Delphi study: attributes of a 'high-quality' trainer

46. The trainer acknowledged when he was unable to explain the manoeuvres he had performed.

This item would be suitable for....

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
DOTS	◯	◯	◯	◯	◯
LETS	◯	◯	◯	◯	◯

Please suggest any modifications/justify your answer (optional)

Delphi study: attributes of a 'high-quality' trainer

9. Attributes of an endoscopy trainer: Exploration

47. The trainer allowed the trainee enough time to carry out the procedure without rushing them.

This item should be included in the...

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
DOTS	◯	◯	◯	◯	◯
LETS	◯	◯	◯	◯	◯

Please suggest any modifications/justify your answer (Optional)

48. The trainer helped the trainee to carry out the procedure independently whilst ensuring patient safety.

This item should be included in the...

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
DOTS	◯	◯	◯	◯	◯
LETS	◯	◯	◯	◯	◯

Please suggest any modifications/justify your answer (Optional)

49. The trainer allowed the trainee to attempt procedures independently learning from what does and does not work as long as patient safety is not compromised.

This item should be included in the...

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
DOTS	◯	◯	◯	◯	◯
LETS	◯	◯	◯	◯	◯

Please suggest any modifications/justify your answer (Optional)

Delphi study: attributes of a 'high-quality' trainer

50. The trainer allowed the trainee to find their limits and understand when they should give up trying.

This item should be included in the...

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
DOTS	◯	◯	◯	◯	◯
LETS	◯	◯	◯	◯	◯

Please suggest any modifications/justify your answer (Optional)

51. The trainer handed back the scope after overcoming a difficult manoeuvre.

This item should be included in the...

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
DOTS	◯	◯	◯	◯	◯
LETS	◯	◯	◯	◯	◯

Please suggest any modifications/justify your answer (Optional)

52. The trainer and the trainee agreed goals for future sessions.

This item should be included in the...

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
DOTS	◯	◯	◯	◯	◯
LETS	◯	◯	◯	◯	◯

Please suggest any modifications/justify your answer (Optional)

Delphi study: attributes of a 'high-quality' trainer

10. Attributes of an endoscopy trainer: Reflection and Feedback

53. The trainer provided feedback close to the teaching event.

This item should be included in the...

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
DOTS	◯	◯	◯	◯	◯
LETS	◯	◯	◯	◯	◯

Please suggest any modifications/justify your answer (Optional)

54. The trainer delivered the feedback in a structure appropriate for the trainee.

This item should be included in the...

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
DOTS	◯	◯	◯	◯	◯
LETS	◯	◯	◯	◯	◯

Please suggest any modifications/justify your answer (Optional)

55. The trainer helped the trainee to reflect on their performance.

This item should be included in the...

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
DOTS	◯	◯	◯	◯	◯
LETS	◯	◯	◯	◯	◯

Please suggest any modifications/justify your answer (Optional)

56. The trainer reinforced positive aspects of the trainee's performance.

This item should be included in the...

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
DOTS	◯	◯	◯	◯	◯
LETS	◯	◯	◯	◯	◯

Please suggest any modifications/justify your answer (Optional)

Delphi study: attributes of a 'high-quality' trainer

57. The trainer identified aspects for the trainee to develop and improve.

This item should be included in the...

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
DOTS	◯	◯	◯	◯	◯
LETS	◯	◯	◯	◯	◯

Please suggest any modifications/justify your answer (Optional)

58. The trainer helped the trainee to assess if the goals for the session had been achieved.

This item should be included in the...

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
DOTS	◯	◯	◯	◯	◯
LETS	◯	◯	◯	◯	◯

Please suggest any modifications/justify your answer (Optional)

59. The trainer used appropriate challenges (e.g. to justify a manoeuvre they have performed) to help the trainee progress.

This item should be included in the...

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
DOTS	◯	◯	◯	◯	◯
LETS	◯	◯	◯	◯	◯

Please suggest any modifications/justify your answer (Optional)

Delphi study: attributes of a 'high-quality' trainer

11. Attributes of an endoscopy trainer: Global

60. The trainer scheduled enough training lists for the trainee.

This item should be included in the...

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
DOTS	◯	◯	◯	◯	◯
LETS	◯	◯	◯	◯	◯

Please suggest any modifications/justify your answer (Optional)

61. The trainer limited the number of trainees he was teaching to ensure that each trainee received adequate training.

This item should be included in the...

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
DOTS	◯	◯	◯	◯	◯
LETS	◯	◯	◯	◯	◯

Please suggest any modifications/justify your answer (Optional)

62. The trainer gets to know the trainee personally e.g. career aims, family etc.

This item should be included in the...

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
DOTS	◯	◯	◯	◯	◯
LETS	◯	◯	◯	◯	◯

Please suggest any modifications/justify your answer (Optional)

Delphi study: attributes of a 'high-quality' trainer

63. The trainer developed a good working relationship with the trainee.

This item should be included in the...

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
DOTS	◯	◯	◯	◯	◯
LETS	◯	◯	◯	◯	◯

Please suggest any modifications/justify your answer (Optional)

64. The trainer and trainee agreed and worked towards common goals during the training period with a long term training plan.

This item should be included in the...

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
DOTS	◯	◯	◯	◯	◯
LETS	◯	◯	◯	◯	◯

Please suggest any modifications/justify your answer (Optional)

65. The trainer set a good professional example through their own behaviour.

This item should be included in the...

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
DOTS	◯	◯	◯	◯	◯
LETS	◯	◯	◯	◯	◯

Please suggest any modifications/justify your answer (Optional)

66. The trainer collected data to use as feedback to the trainee e.g. DOPS form, CuSum etc.

This item should be included in the...

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
DOTS	◯	◯	◯	◯	◯
LETS	◯	◯	◯	◯	◯

Please suggest any modifications/justify your answer (Optional)

Delphi study: attributes of a 'high-quality' trainer

12. Attributes of an endoscopy trainer: Content

67. The trainer taught according to the guidelines as per JAG and the DOPS.

This item should be included in the...

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
DOTS	◯	◯	◯	◯	◯
LETS	◯	◯	◯	◯	◯

Please suggest any modifications/justify your answer (Optional)

68. The trainer can explain the mechanics of endoscopy.

This item should be included in the...

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
DOTS	◯	◯	◯	◯	◯
LETS	◯	◯	◯	◯	◯

Please suggest any modifications/justify your answer (Optional)

69. The trainer demonstrated their competence at performing endoscopy.

This item should be included in the...

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
DOTS	◯	◯	◯	◯	◯
LETS	◯	◯	◯	◯	◯

Please suggest any modifications/justify your answer (Optional)

70. The trainer taught the importance of feel (tactile feedback).

This item should be included in the...

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
DOTS	◯	◯	◯	◯	◯
LETS	◯	◯	◯	◯	◯

Please suggest any modifications/justify your answer (Optional)

Delphi study: attributes of a 'high-quality' trainer

71. The trainer had a broad knowledge about the practice of endoscopy.

This item should be included in the...

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
DOTS	◯	◯	◯	◯	◯
LETS	◯	◯	◯	◯	◯

Please suggest any modifications/justify your answer (Optional)

72. The trainer taught the trainee to thoroughly examine the mucosa

This item should be included in the...

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
DOTS	◯	◯	◯	◯	◯
LETS	◯	◯	◯	◯	◯

Please suggest any modifications/justify your answer (Optional)

73. The trainer taught the trainee to keep the lumen in view.

This item should be included in the...

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
DOTS	◯	◯	◯	◯	◯
LETS	◯	◯	◯	◯	◯

Please suggest any modifications/justify your answer (Optional)

74. The trainer taught the trainee to keep insufflation to a minimum.

This item should be included in the...

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
DOTS	◯	◯	◯	◯	◯
LETS	◯	◯	◯	◯	◯

Please suggest any modifications/justify your answer (Optional)

Delphi study: attributes of a 'high-quality' trainer

75. The trainer taught the whole process of endoscopy to the trainee e.g. the indications, consent and communication.

This item should be included in the...

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
DOTS	◯	◯	◯	◯	◯
LETS	◯	◯	◯	◯	◯

Please suggest any modifications/justify your answer (Optional)

76. The trainer emphasised the importance of fine tip control.

This item should be included in the...

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
DOTS	◯	◯	◯	◯	◯
LETS	◯	◯	◯	◯	◯

Please suggest any modifications/justify your answer (Optional)

77. The trainer encouraged the trainee to communicate appropriately with the nurses.

This item should be included in the...

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
DOTS	◯	◯	◯	◯	◯
LETS	◯	◯	◯	◯	◯

Please suggest any modifications/justify your answer (Optional)

78. The trainer encouraged the trainee to communicate appropriately with the patient.

This item should be included in the...

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
DOTS	◯	◯	◯	◯	◯
LETS	◯	◯	◯	◯	◯

Please suggest any modifications/justify your answer (Optional)

Delphi study: attributes of a 'high-quality' trainer

13. Attributes of an endoscopy trainer: Heuristic strategies

79. The trainer took advantage of opportune moments to teach.

This item should be included in the...

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
DOTS	◯	◯	◯	◯	◯
LETS	◯	◯	◯	◯	◯

Please suggest any modifications/justify your answer (Optional)

80. The trainer taught specific non standard strategies from his own experience when he felt this would help the trainee.

This item should be included in the...

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
DOTS	◯	◯	◯	◯	◯
LETS	◯	◯	◯	◯	◯

Please suggest any modifications/justify your answer (Optional)

81. The trainer taught the theory of endoscopy before each new stage.

This item should be included in the...

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
DOTS	◯	◯	◯	◯	◯
LETS	◯	◯	◯	◯	◯

Please suggest any modifications/justify your answer (Optional)

Delphi study: attributes of a 'high-quality' trainer

82. The trainer made the trainee aware of how the equipment worked.

This item should be included in the...

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
DOTS	◯	◯	◯	◯	◯
LETS	◯	◯	◯	◯	◯

Please suggest any modifications/justify your answer (Optional)

83. The trainer gradually increased the difficulty of the tasks set for the trainee.

This item should be included in the...

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
DOTS	◯	◯	◯	◯	◯
LETS	◯	◯	◯	◯	◯

Please suggest any modifications/justify your answer (Optional)

84. The trainer adhered to the learning plan and reviewed the long term progress of the trainee.

This item should be included in the...

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
DOTS	◯	◯	◯	◯	◯
LETS	◯	◯	◯	◯	◯

Please suggest any modifications/justify your answer (Optional)

85. The trainer knows the learning goals of the trainee and works toward these goals.

This item should be included in the...

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
DOTS	◯	◯	◯	◯	◯
LETS	◯	◯	◯	◯	◯

Please suggest any modifications/justify your answer (Optional)

Delphi study: attributes of a 'high-quality' trainer

14. Attributes of an endoscopy trainer.

86. Prior to sending out this survey we analysed the statements and removed any that did not fit the criteria that makes an attribute suitable for an evaluation toolkit (i.e. measurable, assessable, and generic to all procedures and stage of training)although we acknowledge that these may be important components of endoscopy training. Below is a list of these statements:

- The trainer dealt with any lack of insight in the trainee
- The trainer makes a trainee aware of loop resolution when teaching colonoscopy
- The trainer let the trainee handle the scope outside of the patient before using the scope on a patient
- The trainer provided continued support for his trainee even when the the trainee had reached competence to 'sign-off'

Do you think any of these attributes SHOULD be included?

87. Please suggest any new attributes that you feel have not been covered already in this survey.

Appendix 4 Questionnaire for Round 2 of the Delphi

Delphi Round 2: Attributes of an endoscopy trainer

1. Instructions

Thank you for completing the first round of this Delphi questionnaire; we greatly appreciate your time and effort. After the first round some statements were clearly accepted to either the LETS or DOTS or were excluded; this second round questionnaire just contains the statements which were not clearly allocated to either tool therefore there are only twenty-nine statements to review so should be a much quicker task!

For these remaining statements we would like you to decide whether you think they are most suitable for the DOTS, LETS, both or neither. Just to remind you when making the decisions about these statements we are aiming to produce a formative feedback tool that will inform trainers about their performance as trainers therefore these need to be short. It consists of two parts
- the DOTS (directly observed teaching skills) – this could be completed either by a trainee or peer observing a single list
- the LETS (long-term evaluation of teaching skills) – this will evaluate attributes displayed over a longer period of time for instance at the end of an attachment and can therefore only be completed by a trainee.

Please try and consider what feedback you yourselves would want as a trainer or if you are a trainee what feedback you would want to give your trainer; therefore we will hopefully end up with two effective tools that are both informative but practical to use. All the statements used in this project will form part of an endoscopy training handbook so by excluding them it does not mean they are lost.

Following your feedback we have re-worded some statements, or amalgamated by theme. Each page in this survey portrays a different theme – see headings. At the beginning of each theme we have listed the statements that have already been included from that section in order to inform your decisions about the other statements.

If you want further information about what others thought about the statements to inform your decisions in this round then attached to the email is a summary document with more information about the research process and summaries of both the statistics and comments made in the last round

Thanks once again for your participation

Dr Louise Macdougall

On behalf of
Prof. Roger Barton
Dr Sally Corbett
Dr Mark Welfare
Dr Christopher Wells

N.B. This project has been discussed with and is supported by the Endoscopy Training Leads across the UK

1. Name (this will later be anonymonised and is included to enable us to track responses)

Delphi Round 2: Attributes of an endoscopy trainer

2. Rules and flow of session

From this section; the following statements have been accepted to the:

LETS
A.The trainer matched their approach and pace to the needs of the trainee (needs defined by stage, preferred learning style, level of confidence).

No statements have yet been allocated to the DOTS

2. These statements have not been clearly allocated to a component of the toolkit; please decide if you consider them to be suitable to the LETS, DOTS, both or neither.

	DOTS	LETS	Both	Neither
B. The trainer agreed and applied the ground rules including when to intervene.	◯	◯	◯	◯
C. The trainer agreed a common vocabulary with the trainee.	◯	◯	◯	◯
D. The trainer did not overburden the trainee with too many tasks.	◯	◯	◯	◯

If you have any comments about the above statements please make them here; please do not repeat any comments that you made in the previous round

Page 2

Delphi Round 2: Attributes of an endoscopy trainer

3. Objective setting

From this section the following statements have been accepted to the:

LETS

A. The trainer and trainee agreed and worked towards common objectives during the training period with a long term training plan.

B. The trainer reviewed the long term progress of the trainee.

No statements have yet been allocated to the DOTS.

3. The below statements are also in this section but have not been clearly allocated to either part of the toolkit; please indicate whether you think they should be in the DOTS, LETS, both or neither.

	DOTS	LETS	Both	Neither
C. The trainer agreed objectives for the session.	◯	◯	◯	◯

If you have any comments about the above statements please make them here; please do not repeat any comments that you made in the previous round

Page 3

Delphi Round 2: Attributes of an endoscopy trainer

4. Intervention/ Observation

From this section;

No statements have yet been allocated to the LETS.

DOTS

A. The trainer intervened in a timely fashion if the trainee was failing to make progress (either at a predefined time or if the trainee is struggling excessively).

B. The trainer allowed the trainee reasonable time to carry out the procedure.

4. The below statements are also in this section but have not been clearly allocated to either part of the toolkit; please indicate whether you think they should be in the DOTS, LETS, both or neither.

	DOTS	LETS	Both	Neither
C. The trainer's attention to each moment of the procedure was appropriate to the trainee's needs.	◯	◯	◯	◯

If you have any comments about the above statements please make them here; please do not repeat any comments that you made in the previous round

Page 4

Delphi Round 2: Attributes of an endoscopy trainer

5. Teaching Strategies

From this section the following statements have been accepted to the;

LETS

A. The trainer took advantage of opportune moments to teach.

No statements have currently been allocated to the DOTS.

5. The below statements are also in this section but have not been clearly allocated to either part of the toolkit; please indicate whether you think they should be in the DOTS, LETS, both or neither.

	DOTS	LETS	Both	Neither
B. The trainer used teaching aids that can support learning (e.g. the magnetic imager, diagrams, models etc).	◯	◯	◯	◯
C. The trainer advised the trainee to move position to help them achieve the desired outcome.	◯	◯	◯	◯

If you have any comments about the above statements please make them here; please do not repeat any comments that you made in the previous round

Page 5

Delphi Round 2: Attributes of an endoscopy trainer

6. Patient safety and comfort

No statements have been automatically allocated in this section

6. The below statements are also in this section but have not been clearly allocated to either part of the toolkit; please indicate whether you think they should be in the DOTS, LETS, both or neither.

	DOTS	LETS	Both	Neither
A. The trainer always ensured the patient was comfortable and safe and their dignity was maintained	◯	◯	◯	◯
B. The trainer encouraged the trainee to communicate appropriately with the patient.	◯	◯	◯	◯
C. The trainer helped the trainee to understand and correct errors they had made	◯	◯	◯	◯

If you have any comments about the above statements please make them here; please do not repeat any comments that you made in the previous round

Page 6

Delphi Round 2: Attributes of an endoscopy trainer

7. Technical teaching

From this section the following statements have been accepted to the;

LETS

A. The trainer taught the whole process of endoscopy to the trainee e.g. the indications, consent, communication and sedation

DOTS

B. The trainer taught the trainee to keep the lumen in view.

7. The below statements are also in this section but have not been clearly allocated to either part of the toolkit; please indicate whether you think they should be in the DOTS, LETS, both or neither.

	DOTS	LETS	Both	Neither
C. The trainer taught the trainee to thoroughly examine the mucosa.	◯	◯	◯	◯
D. The trainer emphasised the importance of fine tip control.	◯	◯	◯	◯
E. The trainer taught the trainee to maintain appropriate insufflation.	◯	◯	◯	◯
F. The trainer used their knowledge of the interaction between the scope and the anatomy to inform their training e.g. loop resolution.	◯	◯	◯	◯
G. The trainer ensured the trainee produced accurate, comprehensive and easily understood reports.	◯	◯	◯	◯

If you have any comments about the above statements please make them here; please do not repeat any comments that you made in the previous round

Delphi Round 2: Attributes of an endoscopy trainer

8. Feedback and reflection

From this section the following statements have been accepted to the;

LETS

A. The trainer reviews the data collected by the trainee to inform feedback e.g. DOTS form, CuSum etc

DOTS:

B. The trainer helped the trainee to assess if the objectives for the session had been achieved.

8. The below statements are also in this section but have not been clearly allocated to either part of the toolkit; please indicate whether you think they should be in the DOTS, LETS, both or neither.

	DOTS	LETS	Both	Neither
C. The trainer reinforced positive aspects of the trainee's performance.	◯	◯	◯	◯
D. The trainer identified aspects for the trainee to develop and improve.	◯	◯	◯	◯
E. The trainer helped the trainee to reflect on their performance.	◯	◯	◯	◯

If you have any comments about the above statements please make them here; please do not repeat any comments that you made in the previous round

Delphi Round 2: Attributes of an endoscopy trainer

9. Competence/ Professionalism

From this section the following statements have been accepted to the;

LETS

A. The trainer set a good professional example through their own behaviour.

Delphi Round 2: Attributes of an endoscopy trainer

10. Team

In this section;

No statements have yet been allocated to the LETS.

DOTS

A. The trainer ensures the trainee knows the name and the role of each member of the endoscopy team before a training encounter so that the trainee is supported

9. The below statements are also in this section but have not been clearly allocated to either part of the toolkit; please indicate whether you think they should be in the DOTS, LETS, both or neither.

	DOTS	LETS	Both	Neither
B. The trainer encouraged the trainee to communicate appropriately with the nurses.	◯	◯	◯	◯

If you have any comments about the above statements please make them here; please do not repeat any comments that you made in the previous round

Delphi Round 2: Attributes of an endoscopy trainer

11. Interpersonal

From this section the following statements have been accepted to the;

LETS

A. The trainer was available and focussed on the trainee – by minimising distractions

B. The trainer built the trainee's confidence

No statements have yet been allocated to the DOTS.

10. The below statements are also in this section but have not been clearly allocated to either part of the toolkit; please indicate whether you think they should be in the DOTS, LETS, both or neither.

	DOTS	LETS	Both	Neither
C. It was easy to ask the trainer questions	◯	◯	◯	◯
D. The trainer developed a good working relationship with the trainee	◯	◯	◯	◯
E. The trainer made the trainee feel welcome	◯	◯	◯	◯
F. The trainer was patient and calm	◯	◯	◯	◯

If you have any comments about the above statements please make them here; please do not repeat any comments that you made in the previous round

Page 11

Delphi Round 2: Attributes of an endoscopy trainer

12. Trainee's articulation

No statements have yet been allocated in this section.

11. The below statements are also in this section but have not been clearly allocated to either part of the toolkit; please indicate whether you think they should be in the DOTS, LETS, both or neither.

	DOTS	LETS	Both	Neither
A. The trainer actively listened to the trainee	◯	◯	◯	◯
B. The trainer checks that the trainee has understood what has been said through observation and direct questioning	◯	◯	◯	◯
C. The trainer questioned the trainee at appropriate times.	◯	◯	◯	◯
D. The trainer asked the trainee to show where they are struggling	◯	◯	◯	◯
E. The quantity of dialogue was appropriate for the trainee and the specific teaching episode	◯	◯	◯	◯

If you have any comments about the above statements please make them here; please do not repeat any comments that you made in the previous round

Page 12

Delphi Round 2: Attributes of an endoscopy trainer

13. Trainer scaffolding

From this section;

No statements have yet been allocated to the LETS.

DOTS

A. The trainer provided explanations and descriptions at appropriate times

B. The trainer used a mixture of suggestions, prompts, solutions and instructions to the trainee

C. The trainer demonstrated a procedure to the trainee where necessary

12. The below statements are also in this section but have not been clearly allocated to either part of the toolkit; please indicate whether you think they should be in the DOTS, LETS, both or neither.

	DOTS	LETS	Both	Neither
D. The trainer reassured the trainee at appropriate times.	◯	◯	◯	◯
E.The trainer helped the trainee to carry out the procedure independently	◯	◯	◯	◯

If you have any comments about the above statements please make them here; please do not repeat any comments that you made in the previous round

Appendix 5 Summary of Delphi Round 1 Responses Sent to Participants

Summary of Delphi Round 1: Devising a toolkit to evaluate the endoscopy trainer

---

Thank you for completing the first round of this Delphi questionnaire; we greatly appreciate the time and effort you put in. We have analysed the results and the vast majority of statements were felt suitable for the toolkit; we have however tried to take on board all the comments you have made. This document summarises the statistics and comments for each item so that you can see what your colleagues thought of the statements and where you and your colleagues allocated them; this should help inform your decisions for this round.

Analysis

Some of the statements were clearly felt to be suitable for either the DOTS or LETS and have therefore been allocated to this part of the toolkit; we have still reviewed your comments and amalgamated statements or changed the wording at your suggestion. However we will not ask you to look at these again in this round. Some statements did not meet our consensus threshold; we have reviewed these statements and they have been excluded unless it was due to miscomprehension in which case they have been reworded and are for review here.

The remainder of the statements met the consensus threshold but were not clearly allocated to either tool; it is these statements that we wish you to review. Some comments occurred frequently across the statements; these were that some of the statements were

- too similar and required amalgamation
- not measurable/rateable by the trainee
- not the trainer's responsibility
- poorly worded or alternative wording suggested
- too generic and statements needed to be more specific
- mutually exclusive

As some of these comments consistently reappeared across the statements we have regrouped the statements into more obvious themes and amalgamated statements where appropriate; the rest of the above list we have used as a generic set of criteria alongside each statements individual comments to enable us to modify or exclude statements where required.

Statement summaries

Below are the statements listed within the new categories; for each statement there is an explanation of any modifications and a summary of the comments; we have not included statements that have been excluded.

2. Rules and flow of the session

A. The trainer matched their approach and pace to the needs of the trainee (needs defined by stage, preferred learning style, level of confidence)					
					Allocated to LETS.
DOTS	Total Agree 76.1%	Strongly agree 42.3%	LETS	Total Agree 43.7%	Strongly Agree 83.1%
Wording altered from '*Q 28. The style and pace of the training was appropriate to the trainee.'*					

B. The trainer agreed and applied the ground rules including when to intervene					
					For review
DOTS	Total Agree 83.1%	Strongly agree 33.8%	LETS	Total Agree 84.5%	Strongly Agree 32.4%
Wording adapted from '*Q5. The trainer agreed the rules of the training and was consistent in the application of these rules'.*					
Summary of comments • Important for the entire programme of training • Rules should be clear and explicit • Suggest using word ground rules • Statement discourages flexibility in training • Occur at the start of training but does not require re-iteration					

C. The trainer agreed a common vocabulary with the trainee.					
					For review
DOTS	Total Agree 70.4%	Strongly agree 33.8%	LETS	Total Agree 69%	Strongly Agree 33.8%
No change to wording					
Summary of comments • Doesn't need to happen every time • 'Effective communication needed but this is excessive'					

D. The trainer did not overburden the trainee with too many tasks					
					For review
DOTS	Total Agree 71.8%	Strongly agree 35.2%	LETS	Total Agree 64.8%	Strongly Agree 31%
Wording adapted from *'Q20. The trainer concentrated on one thing at a time and did not overburden the trainee.'*					
Summary of comments • Dependent on level of trainee • Not worth asking for every list • Agree with not overburdening the trainee but not 'one thing at a time'; trainees are often expected to be able to handle more than one task					

- Important

3. Setting Objectives

One of the general comments about this section was that the word objective is more commonly used therefore the wording has been changed to reflect this.

A. The trainer and trainee agreed and worked towards common objectives during the training period with a long term training plan.					
					Allocated to LETS
DOTS Q59	Total Agree 59.1%	Strongly agree 22.6%	LETS Q59	Total Agree 50.7%	Strongly Agree 85.9%
DOTS Q47	Total Agree 88.7%	Strongly agree 50.7%	LETS Q47	Total Agree 47.9%	Strongly Agree 85.9%
Amalgamated from '*Q59 The trainer and trainee agreed and worked towards common objectives during the training period with a long term training plan*' and '*Q47 The trainer and the trainee agreed goals for future sessions*.'					

B. The trainer reviewed the long term progress of the trainee					
					Allocated to LETS.
DOTS Q79	Total Agree 54.9%	Strongly agree 15.5%	LETS Q79	Total Agree 84.5%	Strongly Agree 26.8%
DOTS Q80	Total Agree 74.6%	Strongly agree 23.9%	LETS Q80	Total Agree 85.5%	Strongly Agree 33.8%
Statement an amalgamation of '*Q79. The trainer adhered to the learning plan and reviewed the long term progress of the trainee*' and '*Q80. The trainer knows the learning goals of the trainee and works toward these goals*.'					

C. The trainer agreed objectives for the session					
					For review
DOTS	Total Agree 87.3%	Strongly agree 36.6%	LETS	Total Agree 78.9%	Strongly Agree 28.2%
Wording adapted from '*Q4. The trainer agreed SMART objectives for the session with his trainee at the start of the list. (S = specific, M = measurable, A = achievable, R = relevant, T = can be achieved in the timeframe).*'					
Summary of comments • Not achievable in every list as often training opportunistic therefore goals should be more medium/long term • There should be an overall plan that does not necessarily need re-stated at the beginning of every list • Only appropriate for trainees in the early stages of training • Before the list too prescriptive; might be best done after the first case; at the end of a list					

4. Intervention/ Observation

A. The trainer intervened in a timely fashion if the trainee was failing to make progress (either at a predefined time or if the trainee is struggling excessively).					
					Allocated to DOTS.
DOTS	Total Agree 90.1%	Strongly agree 52.1%	LETS	Total Agree 81.6%	Strongly Agree 36.6%
No changes made					

B. The trainer allowed the trainee reasonable time to carry out the procedure					
					Allocated to DOTS.
DOTS	Total Agree 84.5%	Strongly agree 28.2%	LETS	Total Agree 71.8%	Strongly Agree 18.3%
Wording adapted from *'Q42 The trainer allowed the trainee enough time to carry out the procedure without rushing them.'*					

C. The trainer's attention to each moment of the procedure was appropriate to the trainee's needs					
					For review
DOTS Q24	Total Agree 87.3%	Strongly agree 53.5%	LETS Q24	Total Agree 78.9%	Strongly Agree 45.1%
DOTS Q29	Total Agree 57.7%	Strongly agree 23.9%	LETS Q29	Total Agree 57.7%	Strongly Agree 21.1%
Amalgamation of '*Q24. The trainer closely observed the process and was aware of what the trainee was doing'* and *'Q29 The trainer adjusted the position he was standing in the room appropriately, withdrawing as the trainee progressed.'*					
Summary of comments for both statements • Level of trainee dependent • Not useful, 'by definition where training is occurring there will be observation' • Vague measurement • Should never withdraw completely • Trainer 'used to leave the room once the scope was out the patient… this helped build confidence' • 'Trainers have different styles of teaching including proximity to trainee,					

5. Teaching strategies

A. The trainer took advantage of opportune moments to teach.					
					Allocated to LETS.
DOTS	Total Agree 64.8%	Strongly agree 22.5%	LETS	Total Agree 76.1%	Strongly Agree 25.4%
No changes made					

B. The trainer used teaching aids that can support learning (e.g. the magnetic imager, diagrams, models etc).					
					For review
DOTS	Total Agree 74.7%	Strongly agree 29.6%	LETS	Total Agree 78.8%	Strongly Agree 38%
No changes to wording made.					
Summary of comments • Many felt that this just referred to the magnetic imager and therefore felt that this may lead to disparity as different units have different facilities • Part of efficient teaching but not necessary at every list therefore better for the LETS • Important • Useful but not essential					

C. The trainer advised the trainee to move position to help them achieve the desired outcome					
					For review
DOTS	Total Agree 30.9%	Strongly agree 7%	LETS	Total Agree 23.9%	Strongly Agree 5.6%
Adapted from '*Q25. The trainer physically moved his trainee to help them to achieve the desired outcome*' as many felt it was not always appropriate to physically touch the trainee					
Summary of comments • 'Physically touching generally not a good idea • May be better to advise a move • Judging someone on this may not be appropriate • Did not understand the statement					

6. Patient Safety and Comfort

A The trainer always ensured that the patient was comfortable and safe and their dignity was maintained					
					For review.
DOTS	Total Agree 81.7%	Strongly agree 63.4%	LETS	Total Agree 73.2%	Strongly Agree 53.5%
No changes to wording made					
Summary of comments • Trainee's responsibility or a partnership between trainer and trainee					

- Trainee not in a position to comment
- Difficult for trainee to say no to.
- Useful as a measure during individual cases

B. The trainer encouraged the trainee to communicate appropriately with the patient.					
				For review.	
DOTS	Total Agree 88.7%	Strongly agree 47.9%	LETS	Total Agree 84.5%	Strongly Agree 49.3%
No changes made to wording					
Summary of comments • Fundamental to training					

C. The trainer helped the trainee to understand and correct errors they had made					
				For review.	
DOTS	Total Agree 71.9%	Strongly agree 29.6%	LETS	Total Agree 64.8%	Strongly Agree 25.4%
Adapted from '*Q22. The trainer dealt with any slips, errors or mistakes made by the trainee'.*					
Summary of comments • Trainee should deal with/ be made aware of/ given opportunity to correct • Only deal with if necessary • Should be included in feedback section • Not specific • Some errors/ problems can be allowed to ride; otherwise may hinder progress of case • Criticism of terminology - 'dealt with'; 'slips'					

7. Technical teaching

A. The trainer taught the whole process of endoscopy to the trainee e.g. the indications, consent, communication and sedation					
				Allocated to LETS.	
DOTS	Total Agree 70.5%	Strongly agree 42.3%	LETS	Total Agree 81.7%	Strongly Agree 49.3%
Only change is to add *'and sedation'* as one of the examples.					

B. The trainer taught the trainee to keep the lumen in view.					
				Allocated to DOTS.	
DOTS	Total Agree 84.5%	Strongly agree 47.9%	LETS	Total Agree 76%	Strongly Agree 39.4%
No change to wording.					

C. The trainer taught the trainee to thoroughly examine the mucosa

					For review.
DOTS	Total Agree 87.3%	Strongly agree 49.3%	LETS	Total Agree 78.9%	Strongly Agree 46.5%
No change to wording.					
Summary of comments • Most important • Vague • 'Question too dependent on what the trainee perceives of his own importance'					

D. The trainer emphasised the importance of fine tip control					
					For review.
DOTS	Total Agree 77.5%	Strongly agree 45.1%	LETS	Total Agree 71.8%	Strongly Agree 35.2%
No change to wording					
Summary of comments • Vague wording • More LETS					

E. The trainer taught the trainee to maintain appropriate insufflation					
					For review.
DOTS	Total Agree 78.9%	Strongly agree 35.2%	LETS	Total Agree 71.8%	Strongly Agree 31%
Adapted from '*Q69. The trainer taught the trainee to keep insufflation to a minimum*'					
Summary of comments • Depends on site and circumstance • 'Question too dependent on what the trainee perceives of his own performance' • Using a minimum is sometimes wrong (e.g. on withdrawal)					

F. The trainer used their knowledge of the interaction between the scope and the anatomy to inform their training e.g. loop resolution					
					For review
DOTS	Total Agree 67.6%	Strongly agree 26.8%	LETS	Total Agree 64.8%	Strongly Agree 28.2%
Wording adapted from '*Q63. The trainer can explain the mechanics of endoscopy'*					
Summary of comments • Unsure what is meant by mechanics					

G. The trainer ensured the trainee produced accurate, comprehensive and easily understood reports.
For review
This statement has been added after several suggestions that this should be included as a statement in the previous round.

8. Feedback and reflection

A. The trainer reviews the data collected by the trainee to inform feedback e.g. DOTS form, CuSum etc					
				Allocated to LETS.	
DOTS	Total Agree 69.1%	Strongly agree 25.4%	LETS	Total Agree 78.9%	Strongly Agree 36.6%
Wording adapted from '*Q61. The trainer collected data to use as feedback to the trainee e.g. DOPS form, CuSum etc*.'					

B. The trainer helped the trainee to assess if the objectives for the session had been achieved.					
				Allocated to DOTS.	
DOTS	Total Agree 78.9%	Strongly agree 25.4%	LETS	Total Agree 57.7%	Strongly Agree 18.3%
No changes made to wording.					

C. The trainer reinforced positive aspects of the trainee's performance.					
				For review	
DOTS	Total Agree 83.1%	Strongly agree 35.2%	LETS	Total Agree 78.9%	Strongly Agree 33.8%
No changes made to this statement but '*Q49. The trainer delivered the feedback in a structure appropriate for the trainee*' excluded as implicit in this statement.					
Summary of comments for both statements • Vague • Covered in previous questions • Part of feedback					

D. The trainer identified aspects for the trainee to develop and improve.					
				For review	
DOTS	Total Agree 87.4%	Strongly agree 45.1%	LETS	Total Agree 85.9%	Strongly Agree 46.5%
As with the above statement no changes made to this statement but '*Q49. The trainer delivered the feedback in a structure appropriate for the trainee*' excluded as implicit in this statement					
Summary of comments for all three statements • Vague • Too similar to other statements • Vital in early training • Part of feedback					

E. The trainer helped the trainee to reflect on their performance.					
				For review	
DOTS	Total Agree 81.7%	Strongly agree 33.8%	LETS	Total Agree 78.9%	Strongly Agree 35.2%
No changes made to wording.					

Summary of comments • Too similar to other statements • Part of feedback

9. Competence/Professionalism

A. The trainer set a good professional example through their own behaviour.					
					Allocated to LETS
DOTS	Total Agree 74.7%	Strongly agree 46.5%	LETS	Total Agree 81.7%	Strongly Agree 50.7%
No changes made to wording					

10. Team

A. The trainer ensures the trainee knows the name and the role of each member of the endoscopy team before a training encounter so that the trainee is supported					
					Allocated to DOTS.
DOTS	Total Agree 84.5%	Strongly agree 45.1%	LETS	Total Agree 63.4%	Strongly Agree 25.4%
DOTS	Total Agree 87.4%	Strongly agree 59.2%	LETS	Total Agree 63.4%	Strongly Agree 26.8%
An amalgamation of statements '*Q1. The trainer clarifies everyone's role before a training encounter so that each individual knows how they are involved in the training process'* and *'Q2. The nurses are informed it is a training list to ensure they are supportive of the trainee'*					

B. The trainer encouraged the trainee to communicate appropriately with the nurses.					
					For review.
DOTS	Total Agree 84.5%	Strongly agree 43.7%	LETS	Total Agree 86%	Strongly Agree 43.7%
No changes made to wording					
Summary of comments • A vital tool					

11. Interpersonal

In this section it was felt there was a lot of commonality in the statements and that they could be amalgamated. The other criticism is that often this can be quite subjective from the point of view of the trainee and could be seen as a personal criticism. When reviewing them this was a lengthy category with 11 statements; we have therefore tried to keep those that are more measurable and less personal.

A. The trainer was available and focussed on the trainee – by minimising distractions					
				Allocated to LETS.	
DOTS	Total Agree 78.8%	Strongly agree 39.4%	LETS	Total Agree 84.5%	Strongly Agree 46.5%
Wording adapted from '*Q27 The trainer was appropriately available for the trainee.'*					

B. The trainer built the trainee's confidence					
				Allocated to LETS.	
DOTS	Total Agree 77.4%	Strongly agree 39.4%	LETS	Total Agree 88.8%	Strongly Agree 42.3%
Wording adapted from '*Q11. The trainer built the trainee's self-belief and confidence.'*					

C. It was easy to ask the trainer questions					
				For review.	
DOTS	Total Agree 88.8%	Strongly agree 59.2%	LETS	Total Agree 94.3%	Strongly Agree 53.5%
Adapted from '*Q9. The trainer was approachable and it was easy to ask him questions.'*					
Summary of Comments • Subjective • 'Essential' • 'Should support learning environment on all occasions' • Too similar to other statements; may be better as a combined statement					

D. The trainer developed a good working relationship with the trainee					
				For review.	
DOTS	Total Agree 61.9%	Strongly agree 22.5%	LETS	Total Agree 74.6%	Strongly Agree 22.5%
No change to wording made					
Summary of comments • Too similar to other statements					

E. The trainer made the trainee feel welcome					
				For review.	
DOTS	Total Agree 77.5%	Strongly agree 42.3%	LETS	Total Agree 76%	Strongly Agree 35.2%
No change to wording made.					
Summary of comments • Subjective • Wording too personal • Essential • May already know each other well • Often other department staff that make the trainee feel unwelcome not the trainer					

• Too similar to other statements

F. The trainer was patient and calm.					
					For review.
DOTS	Total Agree 88.7%	Strongly agree 54.9%	LETS	Total Agree 81.7%	Strongly Agree 52.1%
No change to wording made					
Summary of comments • Subjective • Essential; 'the most challenging aspect of training for me' • 'Energetic trainers can be as effective as calm trainers'					

12. Trainee's articulation

A. The trainer actively listened to the trainee					
					For review.
DOTS	Total Agree 81.7%	Strongly agree 38%	LETS	Total Agree 81.7%	Strongly Agree 33.8%
Wording adapted from '*Q33. The trainer provided opportunities for the trainee to speak and actively listened to the trainee*'					
Summary of comments • Could probably be included with other attitudinal questions					

B. The trainer checks that the trainee has understood what has been said through observation and direct questioning					
					For review.
DOTS	Total Agree 81.7%	Strongly agree 32.4%	LETS	Total Agree 70.4%	Strongly Agree 31%
Wording adapted from '*Q37. The trainer checks that trainee has understood what has been said*'					
Summary of comments • Difficult for the trainee to assess as it will be affected by the trainee's perception of whether he understands • Trainee should seek clarification on the points they don't understand • 'Absolute must for therapeutic procedures; not so sure about diagnostic procedures' • 'Checking through observation rather than direct questioning'					

C. The trainer questioned the trainee at appropriate times.					
					For review.
DOTS	Total Agree 78.9%	Strongly agree 35.2%	LETS	Total Agree 71.9%	Strongly Agree 28.2%
No changes made to wording.					
Summary of comments					

• Ground rules for this need to be agreed since some trainees feel uncomfortable if questioned during the procedure itself

D. The trainer asked the trainee to show where they are struggling					
					For review.
DOTS	Total Agree 52.1%	Strongly agree 14.1%	LETS	Total Agree 46.5%	Strongly Agree 8.5%
Wording adapted from '*Q40. The trainer asked the trainee to demonstrate the problem when appropriate'*					
Summary of comments • Statement's meaning unclear • Explain be better than demonstrate					

E. The quantity of dialogue was appropriate for the trainee and the specific teaching episode.					
					For review
DOTS	Total Agree 76.1%	Strongly agree 25.4%	LETS	Total Agree 69%	Strongly Agree 21.1%
Will require a qualifier of too much/too little					
Summary of comments • Subjective • Should be on a scale to indicate whether any failing was due to an excess or paucity of dialogue • 'Amount of chat varies from trainer to trainer, and as long as it works for the trainer I'm not sure it really matters that much' • Worth including					

13. Trainer scaffolding

A. The trainer provided explanations and descriptions at appropriate times					
					Allocated to DOTS.
DOTS q36	Total Agree 81.7%	Strongly agree 35.2%	LETS q36	Total Agree 73.3%	Strongly Agree 28.2%
DOTS q16	Total Agree 91.6%	Strongly agree 66.2%	LETS Q16	Total Agree 78.8%	Strongly Agree 56.3%
An amalgamation of *'Q36. The trainer provided explanations at appropriate times'* and *'Q16. The trainer was able to describe how he performed any endoscopic manoeuvres to his trainee that was understandable to the trainee and the trainee is left with an appreciation of how to perform the procedure'*					

B. The trainer used a mixture of suggestions, prompts, solutions and instructions to the trainee
Allocated to DOTS.

DOTS	Total Agree 87.3%	Strongly agree 47.9%	LETS	Total Agree 76%	Strongly Agree 38%
Original as above but also included '*where appropriate'*					

C. The trainer demonstrated a procedure to the trainee where necessary.					
					Allocated to DOTS.
DOTS	Total Agree 84.5%	Strongly agree 53.5%	LETS	Total Agree 73.3%	Strongly Agree 42.3%
No changes to wording made.					

.

D. The trainer reassured the trainee at appropriate times.					
					For review.
DOTS	Total Agree 76.1%	Strongly agree 32.4%	LETS	Total Agree 66.2%	Strongly Agree 26.8%
No changes made to wording					
Summary of comments • Subjective • 'Reassurance occasionally inappropriate if the trainee's technique is wrong'					

E. **The trainer helped the trainee to carry out the procedure independently**					
					For review
DOTS	Total Agree 83.1%	Strongly agree 39.4%	LETS	Total Agree 77.5%	Strongly Agree 29.6%
Adapted from '*Q43. The trainer helped the trainee to carry out the procedure independently whilst ensuring patient safety.'*					
Summary of comments • Subjective • Difficult for trainee to assess • More useful for independent trainee • Overlaps with other statements					

Appendix 6 DOTS tool used by peers

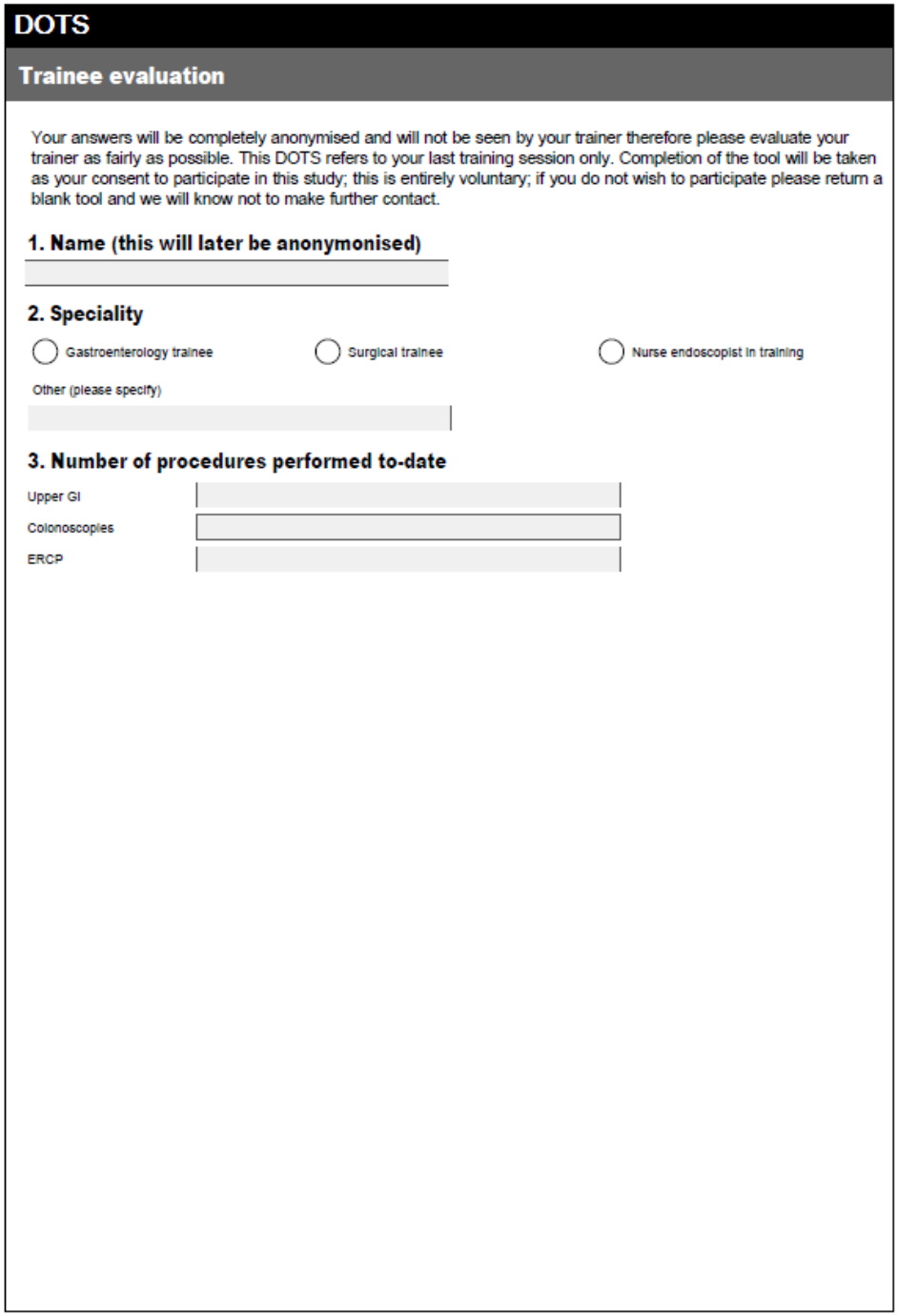
DOTS for peer evaluation

5. The trainer

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
agreed objectives for the session (either previously or at the beginning of the session)	◯	◯	◯	◯	◯
ensured the trainee knew the role and name of each member of the endoscopy team before a training encounter so that they were supported	◯	◯	◯	◯	◯
agreed and applied the ground rules including when to intervene	◯	◯	◯	◯	◯
questioned the trainee at appropriate times	◯	◯	◯	◯	◯
provided explanations and descriptions at appropriate times	◯	◯	◯	◯	◯
used a mixture of suggestions, prompts, solutions and instructions	◯	◯	◯	◯	◯
checked that the trainee had understood instructions and advice by observing or questioning	◯	◯	◯	◯	◯
used an appropriate quantity of dialogue for the trainee and this teaching episode	◯	◯	◯	◯	◯
asked the trainee to show where he/she was struggling	◯	◯	◯	◯	◯
gave specific skills teaching (examples of this might be keeping luminal view, examine the mucosa, tip control, appropriate insufflation, loop resolution)	◯	◯	◯	◯	◯
did not overburden the trainee with too many tasks	◯	◯	◯	◯	◯
demonstrated a procedure where necessary	◯	◯	◯	◯	◯
intervened in a timely fashion (either at a predefined time or if the trainee was struggling)	◯	◯	◯	◯	◯
allowed the trainee reasonable time to carry out the procedure	◯	◯	◯	◯	◯
always ensured that the patient was comfortable and safe and their dignity was maintained	◯	◯	◯	◯	◯
encouraged the trainee to communicate appropriately with the patient	◯	◯	◯	◯	◯
helped the trainee to assess if the objectives for the session had been achieved	◯	◯	◯	◯	◯
reinforced positive aspects of the trainee's performance	◯	◯	◯	◯	◯
identified aspects for the trainee to develop and improve	◯	◯	◯	◯	◯

Comments - please make these comments as specific as possible in order to inform the trainer about their teaching

Appendix 7 DOTS tool used by trainers and trainees



DOTS

Trainee evaluation

Your answers will be completely anonymised and will not be seen by your trainer therefore please evaluate your trainer as fairly as possible. This DOTS refers to your last training session only. Completion of the tool will be taken as your consent to participate in this study; this is entirely voluntary; if you do not wish to participate please return a blank tool and we will know not to make further contact.

1. Name (this will later be anonymonised)

2. Speciality

◯ Gastroenterology trainee ◯ Surgical trainee ◯ Nurse endoscopist in training

Other (please specify)

3. Number of procedures performed to-date

Upper GI	
Colonoscopies	
ERCP	

DOTS

4. The trainer

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
agreed objectives for the session (either previously or at the beginning of the session)	◯	◯	◯	◯	◯
ensured I knew the role and name of each member of the endoscopy team before a training encounter so that I was supported	◯	◯	◯	◯	◯
agreed and applied the ground rules including when to intervene	◯	◯	◯	◯	◯
questioned me at appropriate times	◯	◯	◯	◯	◯
provided explanations and descriptions at appropriate times	◯	◯	◯	◯	◯
used a mixture of suggestions, prompts, solutions and instructions	◯	◯	◯	◯	◯
checked that I had understood instructions and advice by observing or questioning	◯	◯	◯	◯	◯
used an appropriate quantity of dialogue for me and this teaching episode	◯	◯	◯	◯	◯
asked me to show where I was struggling	◯	◯	◯	◯	◯
gave specific skills teaching (examples of this might be keeping luminal view, examine the mucosa, tip control, appropriate insufflation, loop resolution)	◯	◯	◯	◯	◯
did not overburden me with too many tasks	◯	◯	◯	◯	◯
demonstrated a procedure where necessary	◯	◯	◯	◯	◯
intervened in a timely fashion (either at a predefined time or if I am struggling)	◯	◯	◯	◯	◯
allowed me reasonable time to carry out the procedure	◯	◯	◯	◯	◯
always ensured that the patient was comfortable and safe and their dignity was maintained	◯	◯	◯	◯	◯
encouraged me to communicate appropriately with the patient	◯	◯	◯	◯	◯
helped me to assess if the objectives for the session had been achieved	◯	◯	◯	◯	◯
reinforced positive aspects of my performance	◯	◯	◯	◯	◯
identified aspects for me to develop and improve	◯	◯	◯	◯	◯

Comments - please make these comments as specific as possible in order to inform your trainer about their teaching

DOTS

Self-evaluation for trainer

Thank you for completing this tool, we are collecting self-evaluations to analyse how well they correlate with trainee evaluations. Please remember this tool refers to a single session which has been identified between you and your trainee, please ensure you have given them the code supplied to you so that they are able to also complete a trainee evaluation of the same session.

Completion of the tool will be taken as your consent to participate in this study; this is entirely voluntary; if you do not wish to participate please return a blank tool and we will know not to make further contact.

1. Speciality

◯ Gastroenterologist ◯ Surgeon ◯ Nurse Endoscopist

Other (please specify)

2. Have you attended a JAG 'training the trainers' or other teaching course?

◯ Yes ◯ No

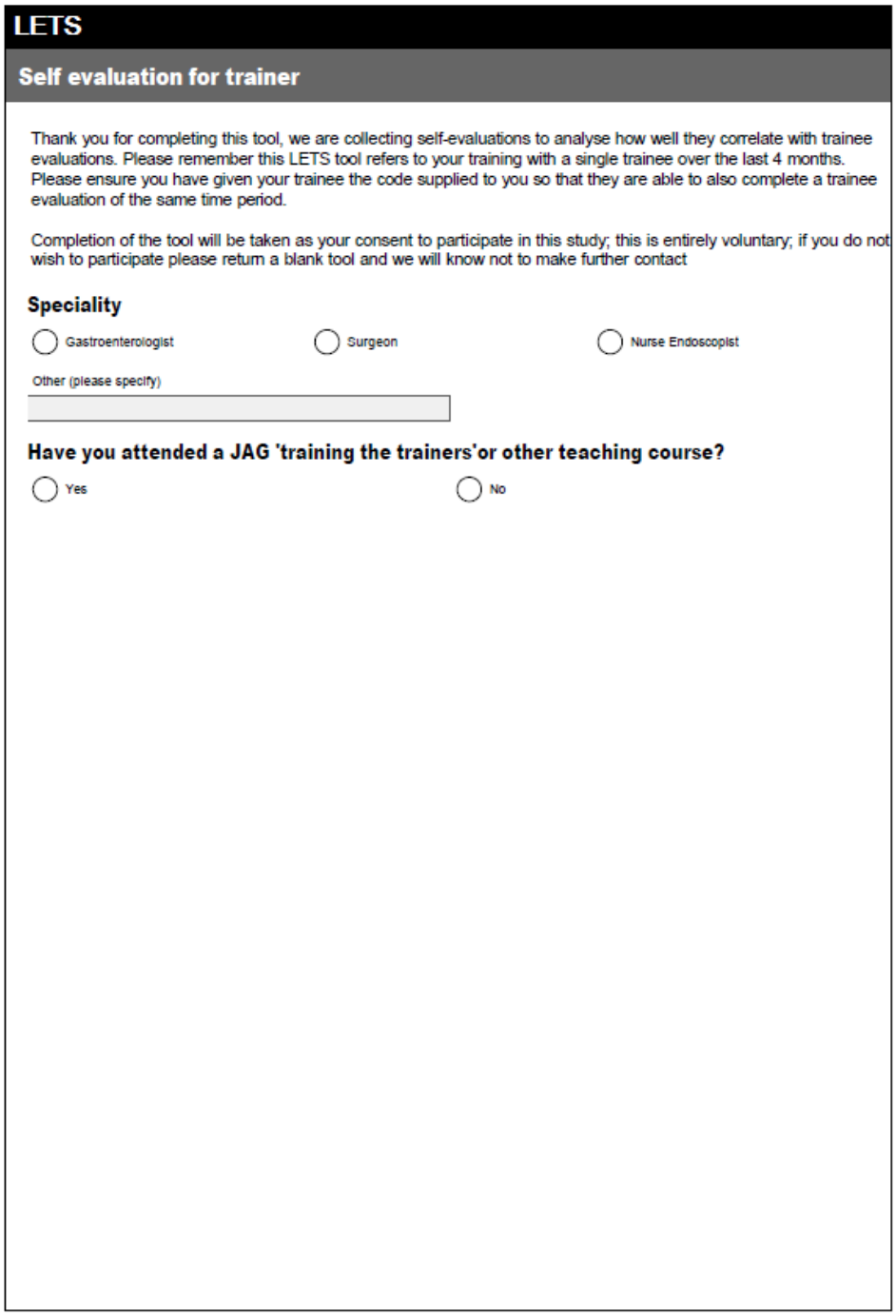
DOTS

3. I, as the trainer,

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
agreed objectives for the session (either previously or at the beginning of the session)	◯	◯	◯	◯	◯
ensured the trainee knew the role and name of each member of the endoscopy team before a training encounter so that they were supported	◯	◯	◯	◯	◯
agreed and applied the ground rules including when to intervene	◯	◯	◯	◯	◯
questioned the trainee at appropriate times	◯	◯	◯	◯	◯
provided explanations and descriptions at appropriate times	◯	◯	◯	◯	◯
used a mixture of suggestions, prompts, solutions and instructions	◯	◯	◯	◯	◯
checked that the trainee had understood instructions and advice	◯	◯	◯	◯	◯
used an appropriate quantity of dialogue for the trainee and this teaching episode	◯	◯	◯	◯	◯
asked the trainee to show me where he/she was struggling	◯	◯	◯	◯	◯
gave specific skills teaching (examples of this might be keeping luminal view, examine the mucosa, tip control, appropriate insufflation, loop resolution)	◯	◯	◯	◯	◯
did not overburden the trainee with too many tasks	◯	◯	◯	◯	◯
demonstrated a procedure where necessary	◯	◯	◯	◯	◯
intervened in a timely fashion (either at a predefined time or if the trainee was struggling)	◯	◯	◯	◯	◯
allowed the trainee reasonable time to carry out the procedure	◯	◯	◯	◯	◯
always ensured that the patient was comfortable and safe and their dignity was maintained	◯	◯	◯	◯	◯
encouraged the trainee to communicate appropriately with the patient	◯	◯	◯	◯	◯
helped the trainee to assess if the objectives for the session had been achieved	◯	◯	◯	◯	◯
reinforced positive aspects of the trainee's performance	◯	◯	◯	◯	◯
identified aspects for the trainee to develop and improve	◯	◯	◯	◯	◯

Comments

Appendix 8 LETS tool including CTEI used by trainers and trainees



LETS

Self evaluation for trainer

Thank you for completing this tool, we are collecting self-evaluations to analyse how well they correlate with trainee evaluations. Please remember this LETS tool refers to your training with a single trainee over the last 4 months. Please ensure you have given your trainee the code supplied to you so that they are able to also complete a trainee evaluation of the same time period.

Completion of the tool will be taken as your consent to participate in this study; this is entirely voluntary; if you do not wish to participate please return a blank tool and we will know not to make further contact

Speciality

◯ Gastroenterologist ◯ Surgeon ◯ Nurse Endoscopist

Other (please specify)

Have you attended a JAG 'training the trainers'or other teaching course?

◯ Yes ◯ No

LETS

I, as the trainer,

	Strongly disagree	Disagree	Neutral	Agree	Strongly Agree
made the trainee feel welcome	◯	◯	◯	◯	◯
agreed and worked towards common objectives during the training period with a long term training plan	◯	◯	◯	◯	◯
matched my approach and pace to the trainee's needs (needs defined by stage, preferred learning style, level of confidence)	◯	◯	◯	◯	◯
used teaching aids that can support learning (e.g. the magnetic imager, diagrams, models etc)	◯	◯	◯	◯	◯
took advantage of opportune moments to teach	◯	◯	◯	◯	◯
checked the trainee's understanding of the theory of endoscopy before each new stage	◯	◯	◯	◯	◯
taught the whole process of endoscopy e.g. the indications, consent, communication and sedation	◯	◯	◯	◯	◯
ensured accurate, comprehensive and easily understood reports were produced	◯	◯	◯	◯	◯
gave the trainee opportunities to ask questions	◯	◯	◯	◯	◯
actively listened to the trainee	◯	◯	◯	◯	◯
was patient and calm	◯	◯	◯	◯	◯
was available and focused on the trainee – by minimising distractions	◯	◯	◯	◯	◯
developed a good working relationship with the trainee	◯	◯	◯	◯	◯
set a good professional example through my behaviour	◯	◯	◯	◯	◯
built the trainee's confidence	◯	◯	◯	◯	◯
reviewed the data collected to inform feedback eg DOPS form, CuSum etc	◯	◯	◯	◯	◯
helped the trainee to reflect on the trainee's performance	◯	◯	◯	◯	◯
reviewed the trainee's long term progress	◯	◯	◯	◯	◯

Comments

LETS

Clinical Teaching Effectiveness Instrument

We would be grateful if you could also complete the Clinical Teaching Effectiveness Instrument (CTEI) (Copeland 2000). This is a short previously validated tool which can be used to evaluate teaching. The reason we have asked you to do this is to enable us to evaluate the degree of correlation between the LETS and the CTEI. This is a recognised method of assessing the validity of a new tool.

If you have already completed this previously during this study you do not need to complete it again. If this is the case please leave it blank and contine to the 'done' button.

Copeland, H. (2000). "Developing and Testing an Instrument to Manage the Effectiveness of Clinical Teaching in an Academic Medical Center." Academic Medicine 75(2): 161.

Please tick the box which most aptly applies to your trainer.

	Never/ Rare	Seldom/ Mediocre	Sometimes/ Good	Often/ Very good	Always/ Superb	N/A
Establishes a good learning environment (approachable, nonthreatening, enthusiastic etc.)	◯	◯	◯	◯	◯	◯
Stimulates me to learn independently	◯	◯	◯	◯	◯	◯
Allows me autonomy appropriate to my level/experience/competence	◯	◯	◯	◯	◯	◯
Organises time to allow for both teaching and care giving	◯	◯	◯	◯	◯	◯
Offers regular feedback (both positive and negative)	◯	◯	◯	◯	◯	◯
Clearly specifies what I am expected ot know and do during this training period	◯	◯	◯	◯	◯	◯
Adjusts teaching to my needs (experience, competence, interest etc.)	◯	◯	◯	◯	◯	◯
Asks questions that promote learning (clarifications, probes, Socratic questions, reflective questions, etc.)	◯	◯	◯	◯	◯	◯
Gives clear explanations/ reasons for opinions, advice, actions etc.	◯	◯	◯	◯	◯	◯
Adjusts teaching to diverse settings (bedside, view box, OR, exam room, microscope etc.)	◯	◯	◯	◯	◯	◯
Coaches me on my clinical/ technical skills	◯	◯	◯	◯	◯	◯
Teaches effective patient and/ or family communication skills	◯	◯	◯	◯	◯	◯
Teaches principles of cost-appropriate care (resource utilisation etc.)	◯	◯	◯	◯	◯	◯