Assessing Decision-Making Skills in Surgery:

Collecting Validity Evidence for the Script Concordance Test

Master of Arts in Education – Health Professions Education

Nada Gawad, MD

Faculty of Education

University of Ottawa

July 2018

Acknowledgements

This research would not have been possible without the support and guidance of many people. First, I would like to thank my two research supervisors, Dr. Timothy Wood and Dr. Isabelle Raîche. The guidance you have both provided me over the past two years has taught me so much and ensured that my work is both grounded in education theory and relevant to the surgical context. You have both also encouraged and supported me in applying the knowledge and skills from my Master's work to other projects, while ensuring that I stay on track with my thesis.

Thank you also to the members of my Thesis Advisory Committee, Dr. Eric Dionne, Dr. David Trumpower, and Dr. Katherine Moreau. Within your areas of expertise, you have all pushed me to consider alternative perspectives, learn new research methods, and explore the intricacies of medical education.

This project could not have been completed without the support of research and project assistants. Lindsay Cowley, you have been involved in every stage of this project from identifying the participants to analyzing the data, and I am so thankful for all your hard work. Lesley Ananny, thank you for your logistic support, particularly with navigating all the administrative tasks that come with a research project.

I could not have undertaken this project without the support of the Department of Innovation in Medical Education, who provided both the funding and the opportunity to work with their research team. I must also thank the 44 staff and resident surgeons who gave me their time to participate in this study. I am so grateful for your support during my research years.

Finally, I am backed by an incredible family for whom I am so thankful. To my parents and step-parents, thank you for the guidance you provided in first deciding to pursue my

Master's, and the support you have offered since then. Mom, you have been guiding me through this every step of the way. To my husband, Eddie, I am sure that you know more about surgical education than you ever thought you would know. From listening to me practice my presentations, to helping me format graphs, to navigating the ups and downs with me, thank you for always having my back as I embarked on my research years.

Abstract

Most in-hospital adverse events are attributable to surgical care and of these, clinical decision-making (CDM) errors account for approximately half. The assessment of CDM is integral to establishing competence among surgical trainees. One proposed assessment tool is the script concordance test (SCT), which is based on the dual process theory of CDM, but evidence demonstrating valid results is needed. This thesis collects content and response process validity evidence for the assessment of CDM using the SCT.

To gather content evidence, a Delphi technique was conducted with a panel of local general surgeons (n=15) consisting of the top decision-makers voted by staff and resident surgeons. Items achieving consensus were mapped on a table of specifications to determine the breadth of topics covered as defined by the Royal College medical expert competencies in general surgery. The final SCT was administered to 29 residents and 14 staff surgeons and results were analyzed. To gather response process evidence, cognitive interviews were then conducted with ten residents and five staff surgeons based on results of the final SCT. Data from the cognitive interviews were analyzed using *a priori* deductive codes based on Tourangeau's cognitive model of response process.

The first round of the Delphi yielded agreement ranging from 40-100% and consensus for 21 cases. The 21 cases made up the final SCT and encompassed 13 of the 19 competencies in general surgery. The final SCT reflected a test of the intraoperative management of open, adult general surgery. Notable absent competencies were described by experts to be outside the scope of general surgery, too difficult for the resident level of training, or presenting an unrealistic intraoperative finding.

Cognitive interviews demonstrated variability in CDM among test-takers. Consistent with the dual process theory, test-takers relied on scripts formed through past experiences, when available, to make decisions. However, test-takers' response process was also influenced by issues with respect to their comprehension, recall, and response matching cognitive steps. Due to issues with response matching in particular, when answering an SCT question test-takers indicating different numerical ratings may have the same rationale.

The Delphi technique, table of specifications, and cognitive interviews provide validity evidence supporting the SCT for assessing CDM of general surgery trainees. Substantial issues with respect to the numerical rating scale suggests further revisions to the test format are required before consideration of its use in summative assessment.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Problem Statement & Overview

Clinical decision-making can be defined as a commitment to a course of action that is intended to yield results that are satisfying for specified individuals (Yates, 2003) and can be interchangeably referred to as judgment or clinical reasoning (Norman, 2005). Ensuring competence in decision-making is thought to be integral to achieving good surgical outcomes (Yule, Flin, Paterson-Brown, Maran, & Rowley, 2006), and thus the assessment of decision-making is integral to establishing competence in surgical training. To illustrate the importance, the majority of in-hospital adverse events are attributable to surgical care (de Vries, Ramrattan, Smorenburg, Gouma, & Boermeester, 2008), and of these, errors in clinical decision-making (CDM) contribute to 30-66% of adverse events (Fabri & Zayas-Castro, 2008; Regenbogen et al., 2007; Rogers et al., 2006).

Given the high prevalence of adverse events due to decision making errors, educators have argued that CDM should be taught and tested by medical schools, licensing bodies, and specialty societies (Norman 2005). Despite this call to include decision-making as part of assessed competencies, little structured assessment of decision-making currently has been implemented in surgical training programs (Cooke & Lemay, 2017). As a key skill developed through residency training, CDM needs to be purposefully and reliably assessed to ensure complete competence (Cooke & Lemay, 2017). Currently how to best measure decision-making skills in this context accurately, reliably, and feasibly is uncertain (Cooke & Lemay, 2017; Groves, Scott, & Alexander, 2002). Excluding the assessment of decision-making skill in determining competence poses a threat to safe surgical care during and after completion of

residency training. Thus, thorough investigation of the assessment of clinical decision-making skills is necessary to establishing competence amongst surgical residency.

In this chapter, I first outline the role of decision-making in surgical training. To understand the overarching theories that inform this thesis, I then delve into decision-making theory, Tourangeau's cognitive model on response process, and modern validity theory. These theories serve as the foundation of understanding CDM assessment tools. Next, I introduce currently used and alternative methods of assessing CDM, including the Script Concordance Test (SCT), which is the CDM assessment tool studied in this thesis. Finally, I review the existing literature on the SCT including its validity evidence, criticisms, and gaps to establish the rationale and research questions for this thesis.

Clinical Decision-Making as a Part of Surgical Competence

Competence is defined as the habitual and judicious use of communication, knowledge, technical skills, clinical decision-making, emotions, values , and reflection in daily practice (Epstein & Hundert, 2002). Errors in surgical practice can be made in a single component of competence independent of the others (Gawande, Zinner, Studdert, & Brennan, 2003). Therefore, experts have argued for the inclusion of decision-making skills in objective structured assessments (Sharma, 2004). The training of medical experts who demonstrate effective decision-making is supported by the CanMEDS Framework that has been integrated into the Royal College's accreditation standards (Royal College of Physicians and Surgeons of Canada, 2005).

This thesis focuses on the CanMEDS' Medical Expert role. From the above definition of competence, the components that fall within CanMEDS' Medical Expert role are knowledge,

technical skills, and CDM (The Royal College of Physicians and Surgeons of Canada, 2018). Knowledge and technical skill are pre-requisites to making sound decisions (Flin, Youngson, & Yule, 2007; Reyna & Adam, 2003), but they are not sufficient. With respect to knowledge, despite the correct comprehension of facts, individuals still may not derive different meaning, which is key to informed decision-making (Reyna, 2008). Similarly, with respect to studies of isolated technical skill versus performance in a complex operating room (OR) environment, there remains a chasm between technical skill proficiency and translation into the operating room because of the dynamic nature of the OR requiring the exercise of appropriate judgment (Aggarwal & Darzi, 2006).

There has been significant work and some success with respect to the assessment of knowledge (Godellas, Hauge, & Huang, 2000) and technical skills (Gofton, Dudek, Wood, Balaa, & Hamstra, 2012; Moorthy, Munz, Sarker, & Darzi, 2003) in surgical training. The purposeful assessment of decision-making skills is a lacking component in determining a surgical trainee's competence within the Medical Expert CanMEDS role (Meterissian, 2006). Further study is necessary to advance valid, reliable, and feasible CDM assessment.

Overarching Theories

In the previous sections I established the need for the assessment of CDM. In this section, I present the overarching theories that relate to this thesis. First, a general overview of modern decision-making theory is introduced along with an introduction specific to script theory, which informs the basis of the method of assessing CDM that is the focus of this thesis. Then Tourangeau's (1984) cognitive model is described as it pertains to respondents' answers to each

item. Finally, I discuss modern validity theory as it relates to establishing the use of any assessment tool.

Decision-making theory.

Although slight variations exist, convention in most scholarship on decision-making defines a decision as a commitment to a course of action that is intended to yield results that are satisfying for specified individuals (Yates, 2003; Yates & Tschirnhart, 2006). Over time, many theories of decision-making have been proposed including those describing analytic processing strategies as well as automatic non-analytic processes (Moulton, Regehr, Mylopoulos, & MacRae, 2007). Modern decision research paradigm focuses on the interfacing of these two processes, and is known as the dual process theory of decision-making. Dual process theory explains how a combination of analytic and non-analytic reasoning are used to solve problems in medical contexts (Brush, Sherbino, & Norman, 2017). The two systems of thinking described in dual process theory are called System 1 and System 2. System 1 thinking is quick and automatic processing and relies on retrieval of recognized patterns and memories (Goos, Schubach, Seifert, & Boeker Martin, 2016; Moulton et al., 2007). System 2 thinking, on the other hand, is slower, analytic and deliberate (Brush et al., 2017; Moulton et al., 2007). Dual process theory proposes that when making a decision, System 1 is subconsciously engaged to retrieve several hypotheses from long-term memory storage and System 2 consciously tests, analyzes, and verifies one of the few hypotheses brought forth (Brush et al., 2017). However, it should be noted that in medicine, and particularly with respect to CDM, the two systems do not represent a dichotomy (Gay & McKinley, 2017; Wood, 2014). Instead, System 1 and System 2 processing are interdependent, although one may predominate (Gay & McKinley, 2017; Wood, 2014). The literature proposes

that several factors may influence which system predominates, including time pressures and the experience of the decision-maker (Brush et al., 2017; Moulton et al., 2007; Pelaccia, Tardif, Triby, & Charlin, 2011).

Experience is one of the fundamental differences between novices and experts. Experiential knowledge is hypothesized to be stored in long-term memory as non-analytic resources known in various literature by several complementary terms, including 'exemplars', 'scripts', and 'schema' which enable rapid retrieval through System 1 thinking (Brush et al., 2017; Moulton et al., 2007; Way et al., 2003). Because novices lack experience, they tend to rely on more analytic reasoning based on causal or conceptual knowledge for early hypothesis generation (Brush et al., 2017). As clinicians gain experience, they become more reliant on these non-analytic resources for early hypothesis generation (Moulton et al., 2007) and the association of a new problem with the clinician's past experience results in improved accuracy and speed of decision-making (Brush et al., 2017). These differences suggest that CDM assessment may focus on the extent to which novices rely on their past experiences when making decisions, as well as the accuracy and speed with which novices make decisions. The reliance on experience to make decisions informs the theory surrounding scripts, which are one of the non-analytic resources used in System 1 processing, and is the foundation of the development of the assessment tool studied in this thesis.

Script theory.

As mentioned, 'scripts' are one variation of the non-analytic resources experts are hypothesized to form through clinical experience. Scripts are memory constructs organized around experience, and form the basis of script theory, introduced in medical education by

Feltovich and Barrows (Feltovich & Barrows, 1984)(Schank & Abelson, 1977) . Script theory takes a prescriptive approach to its application in decision-making, whereby people use existing scripts to understand, interpret, and predict outcomes of new experiences. The actual outcomes of each new experience in turn informs and influences their existing scripts, such that scripts are constantly being updated (Schank, 1982).

Applying script theory to clinical diagnosis, "illness scripts" are developed as novices are exposed to patients and attempt to relate symptoms of real patients to biomedical and clinical knowledge networks they possess to determine a diagnosis or management strategy (Schmidt, Norman, & Boshuizen, 1990). Each encounter with a patient adds to the illness script they form of the constellation of signs and symptoms associated with a given outcome. As novices progress through their training, the hypothesis generation phase of their decision-making process develops from an error-prone, time-consuming, explicit process, to a more efficient and automated script-based process that is characteristic of predominantly System 1 thinking (Charlin, Boshuizen, Custers, & Feltovich, 2007). Illness scripts are then used in the hypothesis-testing phase of the decision-making process to collect, analyze, and interpret data to arrive at a final decision (Dory, Gagnon, Vanpee, & Charlin, 2012).

An individual's repertoire of illness scripts reflects their knowledge and experience, both of which are critical to good decision-making (Goos, Schubach, Seifert, & Boeker Martin, 2016; Stiegler & Gaba, 2015). In medicine, many of the problems faced are charged with uncertainty, ethical challenges, and are ill-defined (Moulton et al., 2007), occupying what has been referred to as the indeterminate zones of practice (Schön, 1987). The use of these scripts represents a cognitive strategy that minimizes cognitive load to enable decision-making in situations of indeterminate zones of practice (Stiegler & Gaba, 2015). As such, the scripts of experts and

novices differ, as does accuracy, quality, and speed of their decision-making (Brush et al., 2017; Goos et al., 2016; Stiegler & Gaba, 2015).

A career in surgery requires daily decision-making that is high-stakes and often in the face of an ambiguous and uncertain clinical picture. Therefore, assessing the decision-making abilities of surgical trainees by investigating their scripts has the potential to assess the trainee's response to indeterminate zones of practice (Dory et al., 2012; Lubarsky et al., 2011), which is integral to providing safe care. Insight into decision-making theory informs the basis of various methods of assessing CDM skills in medical education. The next section focuses on the response process that occurs when answering a question in the context of assessment.

Cognitive model of response process.

An understanding of the cognitive process that occurs during the administration of an assessment tool is essential to ensuring the tool supports its intended use (Padilla & Benitez, 2014). Originally designed for survey response, the cognitive aspects of response process are best espoused by a four-stage model proposed by Tourangeau (Willis, 2015) (Table 1). This model is presented because it is often cited as the dominant source of theoretical framework to support cognitive interviewing (Willis, 2015) as a way of collecting response process validity evidence. Cognitive interviewing is the method employed in this thesis.

Tourangeau's cognitive model (1984) consists of a straightforward sequence of steps. When a question is asked, for example "when is the last time you saw a health professional?", the respondent must first comprehend what the question is asking. The comprehension step pertains to both the specific terms within the question, such as health professional, as well as the overall intent of the question. Next, recall of the relevant information from memory must occur,

or in terms of the example, the respondent must be able to remember their last visit. The third step is the decision-making process, where the respondent must evaluate the information retrieved from memory and make a judgment about when that last visit occurred. The decision-making process may also be influenced by motivation and social desirability, which describes the respondent's desire to respond honestly and thoughtfully or to say something to try and make themselves look better. For example, the respondent may think one year ago may be perceived as too long, and instead make the decision to say six months instead. Finally, response mapping governs how the respondent reports their answer. In this example, the respondent could say 12 months, one year, I'm not exactly sure, or something else. (Willis, 2015)

The decision-making theory presented in the previous section and the cognitive model of response process presented in this section combine to inform the context of responding to CDM questions. For an assessment tool to be used in practice however, the validity of the results generated must be studied. The next section introduces modern validity theory.

Table 1 *Summary of Tourangeau's Four-Stage Cognitive Model of Response Process*

Stage	Description
Comprehension	• Meaning of specific terms and phrases • Understanding of overall question intent
Recall	• Information needed to be recalled to answer the question • Strategies used to recall relevant information
Decision-making	• Mental effort to respond thoughtfully and accurately • Influence of the desire to respond in a way that looks "better"
Response matching	• Matching of internally generated answer with response categories given by the question

Modern validity theory.

For any method of CDM assessment to be considered for use in practice, validity evidence must be established. In this section, I broadly introduce modern validity theory. Validity is "the most fundamental consideration in developing and evaluating tests" (AERA, APA, & NCME, 2014). But in terms of evaluating tests, both classic and modern validity theories agree that validity pertains to the interpretation of data, or the inferences that are made from tests, as opposed to the test or instrument itself (Cronbach, 1971; Kane, 2006). Furthermore, there is a growing consensus toward the unitary view of validity, whereby there are not separate kinds of validity, but rather validity is a single entity for which various sources of evidence are gathered and summarized to arrive at an "integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores" (Messick, 1989).

The *Standards for Educational and Psychological Testing* (AERA et al., 2014) is a set of testing standards that currently represent the gold standard in guidance on testing. Developed jointly by the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME), the five sources of evidence presented in the most recent version include a) test content, b) response processes, c) internal structure, d) relations to other variables, and e) consequences of testing (Table 2).

Test content refers to the themes, wording, and format of the items on an assessment tool (AERA et al., 2014) and is based on analysis of the adequacy with which the test content represents the measured construct and its relevance to the proposed interpretation of test scores (Sireci & Faulkner-Bond, 2014). In other words, content evidence serves to answer "to what

extent does the test cover all possible facets of the construct we are measuring?". The evaluation of test content consists of four elements: domain definition, domain representation, domain relevance, and appropriateness of test construction procedures. Domain definition refers to how the construct being measured by the test is operationally defined by providing details regarding what the test measures. Evaluating domain definition involves achieving external consensus that the operational definition underlying the test is congruent with the domain held by the creators of the test, usually by developing test specifications. Domain representation addresses the degree to which a test represents the targeted domain. Domain relevance refers to the extent to which each item is relevant to the targeted domain. Evaluation of both domain representation and domain relevance is usually done by subject matters experts reviewing items to determine if 1) all important aspects of the content domain are measured by the test, and 2) whether the test contains trivial or irrelevent content. Finally, appropriateness of the test development process addresses quality control procedures during test development for processes including technical accuracy of items, quality of item writing, and susceptibility of items to test-taking strategies. (Sireci & Faulkner-Bond, 2014)

Response process concerns the fit between the construct and the response actually engaged in by examinees or raters (AERA et al., 2014). In other words, response process evidence is the extent to which the thoughts of test takers demonstrate their understanding of the construct as it is defined (Padilla & Benitez, 2014)—"do test-takers understand the questions to mean what we intend them to mean?". Obtaining evidence about response process can be done by one of two methodological categories. The first category is those that directly assess cognitive processes by questioning test takers about their performance strategies or response to particular items, such as in a cognitive interview or focus group (Padilla & Benitez, 2014). The second

category is those that document indirect indicators of performance, and thus require additional inference, such as eye movement or response time (Padilla & Benitez, 2014). Using similar procedures, response process can also be explored by the degree to which raters understand the scoring tasks assigned to them.

Internal structure evidence is defined as the degree to which the relationships among test items conform to the construct on which the proposed test score interpretations are based (AERA et al., 2014; Rios & Wells, 2014). The three main aspects of internal structure are dimensionality, measurement variance, and reliability. Dimensionality refers to determining the extent to which the relationships between items support the intended test scores and their uses. For example, a test that intends to report a single composite score should be unidimensional. Measurement invariance describes evidence that item characteristics are comparable across manifest groups, such as gender and race. Lastly, reliability indices provide evidence that test scores are consistent across multiple administrations. (Rios & Wells, 2014)

Validity evidence in terms of relations to other variables can be expressed in different ways, but pertains to how accurately test scores predict criterion performance, where the criterion represents the outcome of interest. This source of validity evidence allows further determination of the utility of using the test by correlating test scores to external variables. (Oren, Kennet-Cohen, Turvall, & Allalouf, 2014)

Finally, consequences of testing relate to evaluating the soundness of decisions made based on test scores. This applies both in terms of the plausible negative effects as well as the presumed benefits and includes clear statements of the purposes, uses, and outcomes of an assessment system. Specifying the inferences made in the interpretations and uses, evaluating the

proposed inferences using evidence, and considering plausaible alternative interpretations is required. (Lane, 2014)

Table 2 *Modern Validity Framework* (AERA et al., 2014)

Category of Validity Evidence	Definition	Information that could be Reported
Test Content	The degree to which the content represents the construct being measured and its relevance to proposed interpretation of test scores	• How were items chosen • Qualifications of people who chose items • Process for reviewing appropriateness of items
Response Process	The degree to which the thoughts of test takers demonstrate their understanding of the construct as it is defined	• Analysis of learners' thought process • Response time
Internal Structure	The degree to which the relationships among test items conform to the construct on which proposed test score interpretations are based	• Reliability • Item analysis • Factor analysis
Relations to Other Variables	The degree to which scores relate to other measures of similar or dissimilar constructs	• Correlations with other measures
Consequences of Testing	The degree to which sound decisions are made based on test scores, in terms of both positive and negative effects	• Intended and unintended consequences • How pass/fail standard is determined

Summary.

This section provided an introduction to the theory underlying CDM assessment both in terms of how decisions are made as well as a framework that can be used to understand how respondents answer a specific type of question that tests CDM. Then, to ensure test results are meeting the purpose of a test, modern validity theory was introduced. The next section will introduce both currently used and alternatively proposed methods of assessing CDM skills.

The Assessment of Decision-Making Skills

Although fraught with limitations, attempts at assessing CDM currently exist in current surgical training. Other methods of assessing CDM have been proposed and investigated in the literature, but are not widely used in current surgical training. With the understanding of the decision-making theory, cognitive process when answering questions, and validity theory presented in the previous section, this section focuses on the concrete assessment tools that are currently used or alternative proposed for the assessment of CDM.

Currently used methods of assessing decision-making skills.

Despite ample literature on decision-making, it remains an ill-defined construct which makes it difficult to assess (Hall, Ellis, & Hamdorf, 2003). Existing methods of assessment of general surgery trainees in Canada include multiple choice tests, oral examinations, In-Training Evaluation Reports (ITERs), and some form of technical-skills assessment (Sidhu, Grober, Musselman, & Reznick, 2004). Except for oral examinations, the assessment of decision-making skills is not typically the focus of these other assessment tools.

General surgery residents face oral examinations during the RCPSC Certification Exam, and many residency programs also include additional in-training oral exams. These exams can be designed so that decision-making is essential to solving the clinical problems presented (Brailovsky, Charlin, Beausoleil, Coté, & Van Der Vleuten, 2001). There has been some validity evidence consistent with this goal reported (Handfield-Jones, Brown, Rainsberry, & Brailovsky, 1996), but oral examinations have limitations with respect to standardization, reliability of scoring, and administration to large numbers of examinees (Meterissian, 2006; Sidhu et al., 2004) that raises questions as to the value and feasibility of this format in general. Specific

criticisms include their lack of inter-examiner reliability in scoring, scores supporting performance on a narrow scope of clinical cases (and lacking generalizability to other cases), susceptibility to the influence of extraneous factors, and poor correlation with more "objective" measures of knowledge (Schubert, Tetzlaff, Tan, Ryckman, & Mashca, 1999). These weaknesses are exacerbated when administering oral examinations to a large number of trainees, such as in high-stakes certification exams, and this context also poses a feasibility issue (Sidhu et al., 2004).

Alternate methods of assessing decision-making skills.

A variety of other assessment tools have been developed to measure decision-making in postgraduate surgical training, namely chart stimulated recall (CSR), the mini clinical exercise (miniCEX), objective structured clinical examinations (OSCEs), and the key feature tests (KFT). Chart stimulated recall (CSR) has the examinee review several patient charts related to a diagnosis, followed by an interview with an examiner who rates them on several competencies, including decision-making. Issues with CSR typically relate to the costs, resources, and faculty development needed to implement (Holmboe, 2008).

In a miniCEX, faculty observe a learner interacting with a patient in a variety of clinical settings, score the performance of a focused clinical task, and provide feedback. There has been considerable evidence of the benefit of the miniCEX for formative assessment, but summative assessments yield various results due to design, training and scoring issues. The main issues with respect to scoring include leniency of raters and high inter-item correlations (Hawkins, Margolis, Durning, & Norcini, 2010).

OSCEs utilize a hands-on approach to assessing clinical skills (Brailovsky et al., 2001), whereby examinees proceed through a circuit of stations and are presented with a case and a

standardized patient (SP) in each with whom they interact. The station also has as an in-room examiner that scores the examinee using either a checklist, a global-rating scale or both. OSCEs have well-established validity evidence for assessing hands-on clinical skills and decision-making (Brailovsky & Grand'Maison, 2000), but limitations include feasibility of administration, cost, and difficulty separating the evaluation of hands-on clinical skills from the evaluation of decision-making.

The KFT approach is a type of question in which the focus is on the most critical aspect of a case related to patient safety or the most common error that a learner might make when presented with a case. The format of a key feature (KF) question consists of a written patient problem and asks the examinee to select the most pertinent items related to the KF from a list of history, physical examination, and investigation options. With the aim of measuring problem-solving ability, these tests assume there are a limited number of correct responses to a given problem (Page, Bordage, & Allen, 1995).

Finally, the script concordance test (SCT) is a written case-based test designed by Charlin and colleagues (Charlin, Brailovsky, Leduc, & Blouin, 1998). The SCT is founded on the principle that judgments made in the clinical reasoning process can be probed and compared to the judgments made by a panel of experts (Charlin, Roy, Brailovsky, Goulet, & Vleuten, 2000). In this sense, the SCT is a tool designed specifically for assessing clinical decision-making in ambiguous clinical scenarios, which are not well assessed by conventional testing (Fournier, Demeester, & Charlin, 2008). The ability to endorse or eliminate a hypothesis based on a series of judgments is thought to be refined with increasing knowledge and clinical experience (Meterissian, 2006).

The test presents a realistic and challenging clinical case with incorporated uncertainty and an accompanying series of questions that each present several plausible options. An example is provided in Figure 1. Questions are presented in three parts: the first part presents a diagnostic or management option ("if you were thinking of _____") the second part presents a new clinical finding (ex. physical sign, co-morbidity, or lab result), and the third part asks examinees to use a Likert scale to indicate how the clinical finding influences the proposed investigative/management option (B Charlin, Tardif, & Boshuizen, 2000). This Likert scale forces examinees to decide on both the direction (positive, negative, or neutral) and the intensity the finding has on the proposed option (-2, -1, 0, +1, +2).

Case description

You are operating urgently on a 60-year-old female with a complete small bowel obstruction. She has had a previous appendectomy, hysterectomy and sigmoid resection for diverticular disease remotely. You find a tight stricture in the neo-terminal ileum with creeping fat, and thickened mesentery. The remainder of the small bowel is grossly dilated.

If you were planning…	and found…	the planned management is:				
5) ileocecal resection	enlarged nodes in the mesentery	-2	-1	0	+1	+2
6) ileocecal resection	the small bowel perforates as you are mobilizing the right colon	-2	-1	0	+1	+2
7) stricturoplasty	the thickness of the bowel wall is 3 mm	-2	-1	0	+1	+2
8) stricturoplasty	the thickness of the bowel wall is 7 mm for a 2 cm length	-2	-1	0	+1	+2

Investigative action or management

New information

5-anchor Likert-type scale

-2: strongly contraindicated, -1: contraindicated, 0: neither more or less indicated, +1: indicated, +2: strongly indicated

Figure 1. *Example of an SCT Case (Fournier et al, 2008)*

Aggregate scoring is the most commonly used method of scoring as it accounts for the variability of decision-making among experts. Aggregate scoring involves comparing test-takers' answers to a scoring key created from answers provided by a panel of experts. Each expert completes the test independently, and credit for each question is derived from the number of expert panellists who selected each response option divided by the modal value for the question. For example, if for a given question five panel members chose "-2", four chose "-1", and one chose "0", credit for "-2" is 1 (5/5), credit for "-1" is 0.8 (4/5), credit for 0 is 0.2 (1/5), and no credit is given for "+1" or "+2). (Fournier et al., 2008)

Although modifications to the aggregate scoring method exist (Wilson, Pike, & Humbert, 2014), unlike conventional methods of assessing CDM, questions leading to a good level of variability within the expert panel are better for discriminating clinical experience amongst examinees (Charlin et al. 2006), and thus the SCT uniquely does not assume that there is a single best response to a clinical problem. Although weaknesses associated with scoring have been identified (Lineberry, Kreiter, & Bordage, 2013), the strength of this approach is in its design to specifically assess decision-making in situations in which there is uncertainty (Charlin et al., 2000) and its understanding that there may be more than one correct answer (Goos et al., 2016). In summary, there are a wide variety of assessment tools available for measuring decision-making skills. Currently, none are widely accepted as reliable and currently used in assessing decision-making skills in surgery.

The Script Concordance Test

Thus far, I have explained the rationale for the purposeful assessment of CDM as its own component of competence. After introducing the overarching theories relevant to this thesis, I

then briefly introduced various methods of assessing CDM. One method I presented is the SCT. The SCT is the focus of this study because it is designed specifically for assessing CDM in ambiguous scenarios and is based on the premise that there may be more than one correct answer, unlike the other methods of assessing CDM discussed. As such, the SCT is thought to be a step toward the goal of making assessment in medical education more representative of decision-making in practice (Cooke & Lemay, 2017). In this section, I further explore existing literature on the SCT, particularly with respect to existing validity evidence, criticisms of the SCT, and its use in general surgery.

Validity evidence for the SCT.

The SCT supports two of the core principles thought to be paramount to the assessment of surgical decision-making: testing in contexts of uncertainty, and respecting the possibility that there may be more than one acceptable answer (Cooke & Lemay, 2017). As a result, the SCT has been studied across a wide variety of clinical specialties such as surgery (Meterissian, Zabolotny, Gagnon, & Charlin, 2007), emergency medicine (Boulouffe, Doucet, Muschart, Charlin, & Vanpee, 2013), anesthesiology (Ducos et al. 2015), otolaryngology (Iravani, Amini, Doostkam, & Dehbozorgian, 2016), psychiatry (Kazour, Richa, Zoghbi, El-Hage, & Haddad, 2016), radiation oncology (Lambert, Gagnon, Nguyen, & Charlin, 2009), gynecology (Park et al. 2012), and neurology (Tan, Tan, Kandiah, Samarasekera, & Ponnamperuma, 2014). These studies, among others, have made attempts to gather validity evidence for the SCT, which will be discussed by each category of evidence here.

In terms of test content validity, the SCT must present ill-defined and authentic content. Accordingly, Fournier et al. (2008) developed and published guidelines to help test developers

prepare items that meet these needs. Questions must be designed such that factual knowledge is relevant but insufficient, questions are unanswerable using algorithmic reasoning, and cases mimic daily routine (Lubarsky et al., 2011). Further item quality checks recommended pertain to scenarios, questions, and the expert panel. These quality checks are summarized in Table 3.

Table 3 *Script Concordance Test Item Quality Grid (Fournier et al, 2008)*

Scenario	• Describes a scenario that is challenging for experts • Describes an appropriate situation for the population of examinees • Scenario is necessary to set the context and understand the questions • The clinical presentation is typical • The scenario is written correctly (content, grammar)
Questions	• Developed using key-feature approach • Options are relevant as per expert opinion • Same option is not found in two consecutive questions • New information makes it possible to test the link with the investigative action/management for the described case • Answers are spread equally over all the values of the Likert scale • Developed to provide balance between high and low variability
Expert panel	• Consists of 10 to 20 experts • Includes experienced physicians whose presence is appropriate to the level of examinees assessed • Take the test individually, in exactly the same conditions as the examinees

The SCT presumes that examinees mobilize illness scripts from their mental databases, and thus those with more evolved illness scripts make decisions that are more concordant with those of experts (Lubarsky et al., 2011). Across several specialties, studies have demonstrated the SCT's ability to distinguish between medical students, junior residents, senior residents, and/or staff on a statistically significant level, thus supporting this inference (Carrière, Gagnon, Charlin, Downing, & Bordage, 2009; Lambert, Gagnon, Nguyen, & Charlin, 2009; Petrucci,

Nouh, Boutros, Gagnon, & Meterissian, 2013). Empirical evidence to support response process has been demonstrated indirectly through a study demonstrating that examinee's response time was significantly faster and more accurate when presented with information typical of the presented hypothesis than with atypical information (Gagnon et al., 2006). From this study, the authors concluded that processing time and accuracy of judgment on script concordance tasks are affected by how compatible the new clinical information is to relevant activated scripts. Other than this largely theoretical study, there is minimal empirical substantiation to the relationship between the intended construct of the SCT and the thought process of examinees (Lubarsky et al., 2011).

The main source of internal structure evidence for the SCT focuses on reliability, and does not investigate measurement invariance (ex. item discrimination, difficulty) or the assumed unidimensionality (ex. factor analysis). Of the existing reliability evidence, all studies use the same analysis demonstrating high measures of Cronbach's alpha (0.70 to 0.90) as a measure of internal consistency (Iravani, Amini, Doostkam, & Dehbozorgian, 2016; Lambert et al., 2009; Nouh et al., 2012; Park et al., 2010). Additional supportive evidence comes from one study on the expert panel demonstrating that a 10 to 15 member expert panel is acceptable to achieve reliability of alpha > 0.70, but that an expert panel of 20 may be warranted for high-stakes examinations (Gagnon, Charlin, Coletti, Sauve, & van der Vleuten, 2005). Two additional studies, also from the same group, demonstrated that experts directly involved in the training of examinees results in higher absolute scores, but has no bearing on the relative ranking of examinee scores (Charlin, Gagnon, Sauvé, & Coletti, 2007; Sibert et al., 2002).

The relationship of SCT scores to other variables have been investigated in terms of more traditional test formats. One study found a non-significant correlation between a MCQ test and

an SCT (Fournier et al., 2006; Lubarsky et al., 2011), while another study found a positive correlation between a fact-based true/false test and an SCT for lower year students, but not those in later stages of training (Collard et al., 2009). From this, the authors concluded that an independence of factual knowledge and CDM may develop with experience (Collard et al., 2009; Lubarsky et al., 2011). This conclusion may also inadvertently provide some response process evidence by demonstrating CDM developing with experience among test-takers. A SCT written by medical students has also been correlated to other methods of assessing reasoning in later family medicine residency training: simulated office orals ($r=0.447$, $p=0.015$), and short-answer management problems ($r=0.451$, $p=0.013$) (Brailovsky et al., 2001). The correlation was weaker when this SCT was compared to an OSCE designed with the focus of assessing reasoning during hands-on skills ($r=0.352$, $p=0.052$) (Brailovsky et al., 2001). Thus, there is some evidence that the SCT probes a different construct than traditional knowledge-based test formats and demonstrates some overlap in predicting later training scores on tests of similar constructs, but the few relevant studies report modest correlations (Lubarsky et al., 2011).

Finally, consequences of the SCT have been explored by several studies with respect to scoring format, cut-off scores, and impact on learning and teaching practices (Lubarsky et al., 2011). Aggregate scoring has been shown to be a key determinant of the SCT's discriminatory power and has been justified (Charlin et al., 2006; Charlin, Desaulniers, Gagnon, Blouin, & van der Vleuten, 2002), but also has conflicting evidence. Studies have shown similar results using single-best-answer approaches (Bland, Kreiter, & Gordon, 2005) as well as aggregate scoring with distance penalty (Wilson, Pike, & Humbert, 2014). Cut-off scores for determining pass or fail on SCTs has not yet been described (Lubarsky et al., 2011), but a method of transforming scores to provide examinees with insight into their relative performance has been suggested

(Charlin et al., 2010). Interactive workshops for health professionals have used SCT questions to serve as a basis for focused educational discussions between experts and non-experts (Lubarsky et al., 2011). These studies demonstrated validity evidence of self-reported consequences in terms of improvements in participant self-reported knowledge, clinical reasoning skills, and practice habits (Devlin et al., 2008; Labelle et al., 2004; Petrella & Davis, 2007). There is little information on the long-term educational impact of the SCT method on teaching and learning, and no evidence to defend the use of the SCT in high-stakes examinations (Lubarsky et al., 2011).

Criticisms of the SCT.

As described in the previous section, the most commonly used scoring method of the SCT is aggregate scoring, which serves as the basis for most of the issues surrounding the tool (Lineberry, Krieter, & Bordage, 2013). The issues around scoring and reliability are outside the scope of this study, but important to understanding the current limitations of the SCT. As such, in this section I will briefly acknowledge these issues.

While aggregate scoring may offer a novel method of capturing the variation that naturally exists between experts, it also poses unique issues. For example, opposing opinions amongst experts suggesting that contradictory responses are equally correct (Lineberry, Krieter, & Bordage, 2013), examinees possibly outperforming most of the expert panel thus altering the scoring key (Lineberry et al., 2013), and non-extreme modal response questions enabling construct-irrelevant test-taking strategies (Lineberry et al., 2013), may threaten the validity of current scoring approaches. Suggestions to mitigate these issues include discarding items where panellists are uniformly divided on opposing values (Lubarsky, Gagnon, & Charlin, 2013), using

"distance from the mode" scoring where penalty points are a function of the number of steps examinees are away from the modal response (Wilson et al., 2014), purposefully including extreme modal questions to even out response distribution and discourage test-taking strategies, and carefully selecting experts.

Experts are selected as those "who are reputed to have superior knowledge, clinical acumen, and experience in a given domain, and whose opinions are sought (and generally accepted) when challenging cases arise within their field" (Lubarsky, Gagnon, et al., 2013). And while the selection of experts for the SCT is non-evidence-based, empirical medical evidence and expert consensus are also not immune to error (Lubarsky, Gagnon, et al., 2013). Results of previous studies suggest that judicious panellist selection is of utmost importance: one study used test scores from multiple specialties and showed that on a radiation oncology SCT, 27% of residents scored higher than the panellists (Charlin et al., 2010). This was suggested to be because the experts selected were subspecialists within their field, whereas residents harbored undifferentiated knowledge structures (Lubarsky, Gagnon, et al., 2013). On the other hand, the same study used test scores from a SCT in general surgery which strategically chose experts aligned with the content domain of the SCT, and demonstrated that only 1% of residents performed above the panel mean (Charlin et al., 2010). This underscores the importance of expert selection appropriately representing the content domain of the SCT, and suggests that if done well, a significant percentage of examinees should not score above the panel mean.

In terms of reliability, another criticism of the SCT is that inter-panel, inter-panellist, or intra-panellist measurement errors have not been considered (Lineberry et al., 2013). The majority of inferences regarding reliability are determined by Cronbach's alpha coefficient, which is suggested to be an overestimation of true correlation (Schmitt, 1995). The reliability of

the panel as an estimate of true expert opinion (inter-panel reliability) is not an issue in traditional assessments such as conventional multiple-choice exams, where the correct answer is based on expert consensus or empirical evidence (Lineberry et al., 2013). Inter-panellist reliability (or inter-rater reliability) has not been examined extensively, because it is the basis of the aggregate scoring method (Lineberry et al., 2013). But not all panellists can be equally correct, and current scoring methods do not account for the variability between panellists resulting from variable credibility.

SCT in general surgery.

Two studies on the SCT deserve particular mention as they are the only national validation studies in general surgery residents. The first, by Nouh et al. (2012), was administered to 202 residents from PGY1 to PGY5 across nine Canadian universities, and its main strength is its large sample size. The test was created by four program directors, and then four study authors distilled the initial question database down to the best 153 questions based on whether the authors thought each question addressed a realistic intraoperative dilemma and tested CDM skills. Aggregate scoring on a 5 point Likert scale was used. Results demonstrated a reliability (Cronbach alpha) of 0.85, with scores progressively increasing from PGY1 to PGY4 with a dip in the PGY5s. Junior (PGY1 and 2) and Senior (PGY3-5) residents had significantly different scores across all programs. Given the dip in the scores of PGY5s, a follow-up study by Petrucci et al. (2013) was conducted, which hypothesized that graduating PGY5s preparing for their licensing examinations had a broader knowledge of all areas of general surgery compared to the subspecialist expert panel. Thus, they re-scored the SCT using a panel of specialty-specific experts, and found a reliability (Cronbach alpha) of 0.81 and progressively increasing scores

across all levels of training. The main weakness of this pair of studies is that they only evaluated one measure of reliability, Cronbach's alpha coefficient, which may be an overestimation of true correlation as described in the previous section (Schmitt, 1995). Content validity was measured to a limited extent by asking four authors to identify the most suitable questions in terms of whether each question addressed a realistic intraoperative dilemma and tested decision-making skills. The response process of experts versus examinees and the relationship of scores to other methods of assessment were not investigated.

Summary.

In summary, the main criticisms of the SCT include the complexities of aggregate scoring and associated reliability issues, as well as the impact of examinee response style and resulting test-taking strategies. Furthermore, gaps in the existing validity evidence limit the ability to use the SCT in summative assessment. Despite these challenges, the SCT still offers advantages compared to other methods of assessing decision-making skills. The SCT focuses on two concepts considered paramount to the future assessment of decision-making, namely assessment in the context of uncertainty and the acknowledgement that there can be more than one correct answer. This unique focus renders the SCT worthy of further study (Cooke & Lemay, 2017). Furthermore, the two described studies in general surgery trainees have built an SCT upon which further study can be conducted. While addressing the scoring and reliability issues are beyond the scope of this study, addressing the above described gaps in validity evidence while respecting the strengths of the SCT is a necessary next step in reaching the goal of making assessment in medical education more representative of current-day decision making (Cooke & Lemay, 2017).

Literature Review Summary & Purpose Statement

Overall, the literature reveals that there is a disconnect between the importance of assessing decision-making skills and the assessment practices used in surgical residency programs. Furthermore, the only tool that is currently used to assess decision-making is oral examinations both for in-training and certification examinations, but their abilities to measure the decision-making construct is unknown, and their use is impeded by challenges with respect to reliability and feasibility.

Since the SCT was designed specifically to address areas of uncertainty and accepts the possibility of more than one correct answer, it may be a better measure of decision-making in a surgery context than other assessment formats that have been developed and discussed. Its mimicry of real-life challenging clinical scenarios is invaluable to the gap that exists in the assessment of decision-making amongst surgical trainees as it provides an authenticity lacking in more traditional assessment methods. Furthermore, its standardization and ease of administration make it advantageous compared to non-written tests. While considerable work has been done to demonstrate the utility and potential use of the SCT, significant challenges still exist and in-depth studies using all tenets of modern validity theory has yet to be done.

Therefore, the purpose of this study is to explore the use of the SCT as a test of clinical decision-making among general surgery residents at the University of Ottawa for summative purposes. The construct being assessed is clinical decision-making in general surgery, both in acute and elective settings, including the breadth of all its subspecialties. Before implementing any test, it is important to collect validity evidence that demonstrates the test is producing results consistent with its intended use. To that end, modern validity theory (AERA, APA, & NCME,

2014) will be used as a framework to guide the implementation of the SCT. This framework will also serve to ensure the results are as expected and the test is feasible to use. Within this framework, two out of five sources of validity evidence, namely content and response process, will be reported (Table 2).

Feasibility barriers with respect to my Master's degree preclude collecting evidence for all five sources of validity evidence. Test content and response process validity evidence were chosen of the five sources because they serve as the initial steps in gathering validity evidence as they are typically carried out during test development (AERA et al., 2014) while the other three sources of validity evidence (internal structure, relations to other variables, and consequences of testing) are usually collected during pilot testing or after administration of the test. Furthermore, evidence based on test content and response processes are considered complementary and are recommended in the *Standards for Educational and Psychological Testing* for combined use to detect and interpret differential interpretation of the test across groups of examinees (AERA et al., 2014). In addition, the *Standards* specifically recommends using both test content and response process evidence when the rationale for a test use depends on the cognitive processes of test-takers (AERA et al., 2014), such as the presumed assessment of decision-making skills.

Research Questions

Given these gaps and issues, as a preliminary step toward using the SCT to assess CDM among general surgery trainees, this study will investigate two aspects of the validity of the SCT by answering following research questions:

1. Test content validity evidence: To what extent does the SCT used in this study present realistic and challenging clinical general surgery cases that incorporate clinical uncertainty? Does it cover the breadth of general surgery?
2. Response process validity evidence: To what degree do experts and examinees interpret SCT items in a manner consistent with the test's purpose? How do they choose their answer?

Specifically, as shown in Table 2, validity evidence will be determined by investigating the degree to which decision-making is assessed (content validity) as well as the thought processes behind experts' and examinees' responses (response process validity). Of note, the methods as well as results for each of test content and response process validity evidence are described separately in Chapters 2 and 3, respectively.

Chapter 2: Test Content Validity Evidence

Overview

Test content is the first of five sources of validity evidence presented by the *Standards for Educational and Psychological Testing* (AERA et al., 2014). Typically gathered during the test development phase, test content validity evidence represents the measured construct and its relevance to the proposed interpretation of test scores (Sireci & Faulkner-Bond, 2014). Evaluation of test content is usually done by subject matter experts reviewing items to determine if important aspects of the content domain are measured by the test and whether the test contains trivial or irrelevant content (Sireci & Faulkner-Bond, 2014).

A variety of methods involving subject matter experts can be used to determine if the test content is congruent with the SCT's purpose. One of the most common methods involves subject matter experts rating the degree to which items adequately represent their intended content and cognitive specifications and/or the degree to which they are relevant to the domain tested (Sireci & Faulkner-Bond, 2014).

One common strategy in health science education research is a Delphi technique (de Villiers, de Villiers, & Kent, 2005), which was the method used to achieve consensus through expert opinion on the most relevant cases and questions for construction of the SCT for this study. Because the Delphi technique achieves consensus on the individual cases and questions, a commonly used strategy to determine the degree to which the test as a whole represents the content domain is by using a table of specifications as a test blueprint (Sireci, 1988).

Rationale

Delphi.

Delphi methodology is a widely-used process first developed in 1948 by the Rand Corporation, in which participants anonymously express their ideas through multiple iterations of questionnaires (Humphrey-Murto, Varpio, Gonsalves, & Wood, 2017). Responses are collected and analyzed, and the results are provided to participants so that they can reassess and potentially alter their responses from the previous iteration (Cuhls, 2005; Hsu, 2007). As such, the Delphi process is continuously iterated until consensus is achieved (Hsu, 2007). An important advantage of the Delphi over other consensus methods is its feasibility, since experts do not have to physically meet. This improves participation, lowers costs, and allows for experts from different geographic locations and clinical backgrounds to be recruited. In addition, its anonymous nature

ensures that that the group outcome is not skewed by the influence of a dominant group member. (Graham, Regehr, & Wright, 2003)

Rather than create a *de novo* test, an SCT used in a national validation study administered to general surgery residents across Canada (Petrucci et al., 2013) was used for the Delphi. Compared to the original test of 43 cases and 153 questions, literature on the SCT has demonstrated improved reliability in generalizability studies (Fournier et al., 2008) with a test of 20 clinical cases and an average of 3 questions per case. This shortened version is also more feasible with respect to expert and resident participants' completion of the test, and thus we *a priori* planned to achieve consensus in 20 of the 43 cases.

Table of specifications.

The Delphi's rating scale allows determination of how well the items achieve their target, but it may not necessarily provide information about how well general surgery (the domain) is represented in the items achieving consensus (Sireci & Faulkner-Bond, 2014). As such, a table of specifications was used to map the degree to which items represent the targeted content areas (the subspecialties of general surgery). This matching approach can be used to eliminate or revise items, and enables consideration of how the content areas are covered in the test specifications (Fournier et al., 2008; Sireci & Faulkner-Bond, 2014). There are several methods in which a table of specifications can be created. For example, the table of specifications can be evaluated by a panel of expert judges with agreement ratings to achieve consensus (Newman, Lim, & Pineda, 2013). This is typically done for items where there is thought to be variability in expert opinion (Newman et al., 2013), such as relating the item "tomatoes" to the concept of "fruit". Alternatively, when mapping of items is thought to be objective and invariable, a table of specifications can be created by a single subject matter expert, such as a teacher relating the item

"Pythagoras' theorem" to the concept of "Algebra" when creating a mathematics test (Fives & Didonato-Barnes, 2013). For this study, mapping the items to content areas in general surgery does not vary with opinion (i.e. a given item either is or is not a question about the liver), and thus was created by a single subject matter expert with review by a second subject matter expert for any items thought to be ambiguous or related to multiple content areas in general surgery.

A table of specifications can also vary with respect to what concepts the test items are mapped to. For example, a table of specifications can map test items to learning objectives or cognitive levels (Fives & Didonato-Barnes, 2013). Since the extent to which each item measures decision-making may vary with personal opinion, each item's relevance to decision-making was assessed with the Delphi rather than in the table of specifications. As such, the purpose of the table of specifications in this study was to map SCT items with the content areas within general surgery to determine the domain representation aspect of test content validity.

Methods: Test Content Validity

Selection of experts.

An informal survey was distributed to all staff and resident surgeons within the Division of General Surgery at The Ottawa Hospital (TOH) to identify expert decision-makers within the general surgery division. Division members surveyed were asked to select ten staff surgeons they thought to be good decision-makers. Thirty-nine division members responded, and a list encompassing all staff surgeons within the division was formed. The staff surgeons with the highest number of votes from their colleagues and trainees were invited to participate with an email explaining the purpose and methodology of our study. Of the 15 experts initially contacted, 12 experts agreed to participate. An additional three experts were contacted to form a panel of 15

experts. Ten to 15 experts are generally thought to be the minimum number to form a panel (Hsu, 2007), and hence 15 were recruited in case some experts dropped out of the study across rounds. The membership of the expert panel was not revealed to the participants. The campus of practice, clinical area of expertise, and gender varied and are represented in Table 4.

The composition of this sample is roughly comparable to the population of staff general surgeons at their institution. Of note, despite their respective subspecialties, all experts participate in acute care and/or on-call general surgery coverage, thus providing them with additional experience in cases outside their subspecialty of practice.

Table 4 *Composition of the Delphi Expert Panel*

Metric	Expert Panel Composition
Gender	Female: 40% (6/15) Male: 60% (9/15)
The Ottawa Hospital (TOH) Site	General: 53% (8/15) Civic: 47% (7/15)
Clinical Area of Expertise*	HPB: 26.7% (4/15) Breast: 20% (3/15) MIS/Bariatrics: 20% (3/15) Surgical Oncology: 6.7% (1/15) Colorectal: 20% (3/15) Trauma/ACS: 13.3% (2/15)

**Note*. Some surgeons practice in multiple subspecialty areas and thus the total is greater than 100%.

Delphi items.

Delphi items originated from an SCT used in a national validation study administered to general surgery residents across Canada (Petrucci et al., 2013). This test consisted of 43 clinical cases for a total of 153 questions, and had a reliability level (Cronbach α) of 0.81. The only changes made to the test before it was presented to the expert panel were correction of typographical errors.

Delphi methodology.

All expert panellists who agreed to participate were e-mailed a link through which to access access the first round of an on-line survey using a random study number to track their participation. They were informed that the aim of the study was to explore the assessment clinical decision-making skills using the SCT. Instructions on the Delphi process and a description of the SCT's format and structure were provided (Appendices

Appendix 1. Delphi Instructions to Experts), along with a sample case and questions to familiarize experts (Figure 1). Experts were also informed that their responses would be de-identified and feedback consisting of means and standard deviations for each item would be provided prior to re-rating of each item in the subsequent round(s). Each expert was asked to answer five questions about each case:

1. Does it represent a realistic and challenging clinical scenario that incorporates uncertainty?
2. Does it test decision-making skills?
3. Does it relate to a subspecialty of General Surgery defined by the Royal College of Surgeons of Canada?
4. Do you think this is a good question?
5. Why or why not?

Experts were asked to answer the Delphi questions for each SCT case as opposed to each SCT question because of the feasibility limitation with 153 SCT questions. The Delphi questions were constructed based on three of the four elements of content validity described by Sireci (1988), namely domain definition, domain relevance, and appropriateness of test development.

With respect to domain definition, the SCT is operationally defined as a test of decision-making skills that presents realistic and challenging clinical scenarios incorporating uncertainty. Thus, the first two questions were constructed to determine to what extent the expert panel agrees that each item presented does in fact measure this construct. Domain relevance is determined by the third question, which asks if each item relates to a subspecialty of general surgery. Finally, asking if the question is overall considered "good" and allowing for free-text comments broadly investigates the appropriateness of the test development in terms of technical accuracy of items and quality of item writing. Of note, the fourth element, domain representation refers to the test as a whole, and thus was evaluated using the table of specifications as opposed to during the Delphi evaluation of individual items.

In the first round, each expert was asked to answer questions one to four on a Likert scale from one (strongly disagree) to four (strongly agree). A neutral middle point was excluded to encourage experts to choose a side and to enable clear calculations on agreement (de Villiers et al., 2005). To answer question five, participants were given the opportunity to provide written comment on any issues with the case and corresponding questions or otherwise clarify their opinion. This was important to capture the depth of expert opinions, such as why a question was not considered "good", and thus allow improvements to be made. This question also allowed experts to explain if they thought the case was good but some of the questions within it were not. The responses to question five were reviewed to improve or delete each case for the subsequent round. If consensus was not achieved within the first round, subsequent rounds would be performed until consensus is met, to a maximum of three rounds.

Consensus of Delphi data.

Consensus was defined as a condition of homogeneity or consistency within the opinion of expert panellists (Graham et al., 2003). Cases and their corresponding questions were included in the final SCT if both of the following criteria were met:

1. 80% or more of experts rated as 3 (agree) or 4 (strongly agree) for each of the following questions:
    - Does it represent a realistic and challenging clinical scenario that incorporates uncertainty?
    - Does it test decision-making skills?
    - Does it relate to a subspecialty of General Surgery defined by the Royal College of Surgeons of Canada?
2. 80% or more of experts answered 'yes' for the following question:
    - Do you think this is a good question?

Agreement rating was calculated for questions 1-4 for all 43 cases. All statistical analyses were performed using IBM SPSS (Version 24, Chicago, IL, USA).

Table of specifications.

The Delphi was used to cover three of the four elements of content validity: domain definition, domain relevance, and appropriateness of test development. The fourth area described by Sireci (1988) is domain representation, which was determined by constructing a table of specifications. The targeted content areas were defined as the Medical Expert General Surgery Competencies as per the Royal College of Physicians and Surgeons of Canada (RCPSC) (Royal College of Physicians and Surgeons of Canada, 2017). The Royal College guidelines were used

to define the content areas as they represent the standard Canadian trainees must achieve to be licensed for independent practice after they complete residency training. The individual questions, as opposed to the cases, were mapped to capture the multiple competencies covered in each case. Questions were mapped both for the original SCT as well as the final SCT, consisting of only the items achieving consensus after completion of the Delphi. In my role as a subject matter expert, I constructed the table of specifications by assigning each item to one or more competency categories. Any questions that were ambiguous or thought to be represented by multiple competencies were reviewed by a second subject matter expert (Isabelle Raîche).

For all competencies that were included in part of the original SCT but not on the final SCT (i.e. post-Delphi), the written comments provided by the Delphi panel were reviewed to determine the rationale as to why questions pertaining to these competencies were excluded.

Results: Test Content Validity

Delphi

Of the 15 experts recruited, all 15 responded. Percent agreement ranged from 40% to 100% for each question. Expert agreement for all 43 items is listed in Table 5. These results are also depicted graphically in Figure 2. Twenty-one items were identified as reaching consensus, and thus a second round was not needed.

Table 5 *Results of Expert Agreement for SCT Items in Delphi Round 1*

Item Number	Q1: realistic scenario*	Q2: decision making*	Q3: relate to a subspeciatly*	Q4: good question**	Item Included in Final SCT
1	80.0	93.3	93.3	73.3	No
2	100	100	93.3	86.7	Yes
3	93.3	93.3	93.3	86.7	Yes
4	93.3	100.0	93.3	86.7	Yes
5	80.0	93.3	86.7	86.7	Yes
6	86.7	80.0	86.7	46.7	No

7	80.0	80.0	93.3	73.3	No
8	73.3	86.7	60.0	40.0	No
9	73.3	86.7	60.0	40.0	No
10	100.0	93.3	93.3	93.3	Yes
11	73.3	80.0	93.3	73.3	No
12	80.0	86.7	93.3	73.3	No
Item Number	Q1: realistic scenario*	Q2: decision making*	Q3: relate to a subspeciatly*	Q4: good question**	Item Included in Final SCT
13	93.3	93.3	93.3	86.7	Yes
14	93.3	93.3	86.7	66.7	No
15	93.3	93.3	93.3	93.3	Yes
16	73.3	86.7	93.3	53.3	No
17	86.7	93.3	93.3	86.7	Yes
18	93.3	93.3	93.3	86.7	Yes
19	100	93.3	93.3	80.0	Yes
20	93.3	93.3	93.3	93.3	Yes
21	93.3	100.0	93.3	93.3	Yes
22	53.3	73.3	93.3	66.7	No
23	60.0	73.3	80.0	60.0	No
24	80.0	73.3	93.3	66.7	No
25	80.0	93.3	93.3	86.7	Yes
26	80.0	86.7	100.0	80.0	Yes
27	100.0	86.7	93.3	86.7	Yes
28	100.0	93.3	86.7	93.3	Yes
29	86.7	86.7	93.3	86.7	Yes
30	93.3	100.0	93.3	93.3	Yes
31	80.0	86.7	93.3	73.3	No
32	93.3	93.3	80.0	73.3	No
33	93.3	93.3	93.3	80.0	Yes
34	93.3	93.3	80.0	73.3	No
35	80.0	80.0	86.7	66.7	No
36	93.3	100.0	80.0	73.3	No
37	86.7	86.7	73.3	66.7	No
38	93.3	100.0	93.3	93.3	Yes
39	66.7	66.7	86.7	53.3	No
40	60.0	80.0	80.0	60.0	No
41	86.7	86.7	93.3	66.7	No
42	93.3	86.7	100.0	80.0	Yes
43	86.7	93.3	66.7	40.0	No

* = percentage of respondents who gave a rating of ‘agree’ (3) or ‘strongly agree’ (4)
** = percentage of respondents who answered ‘yes’

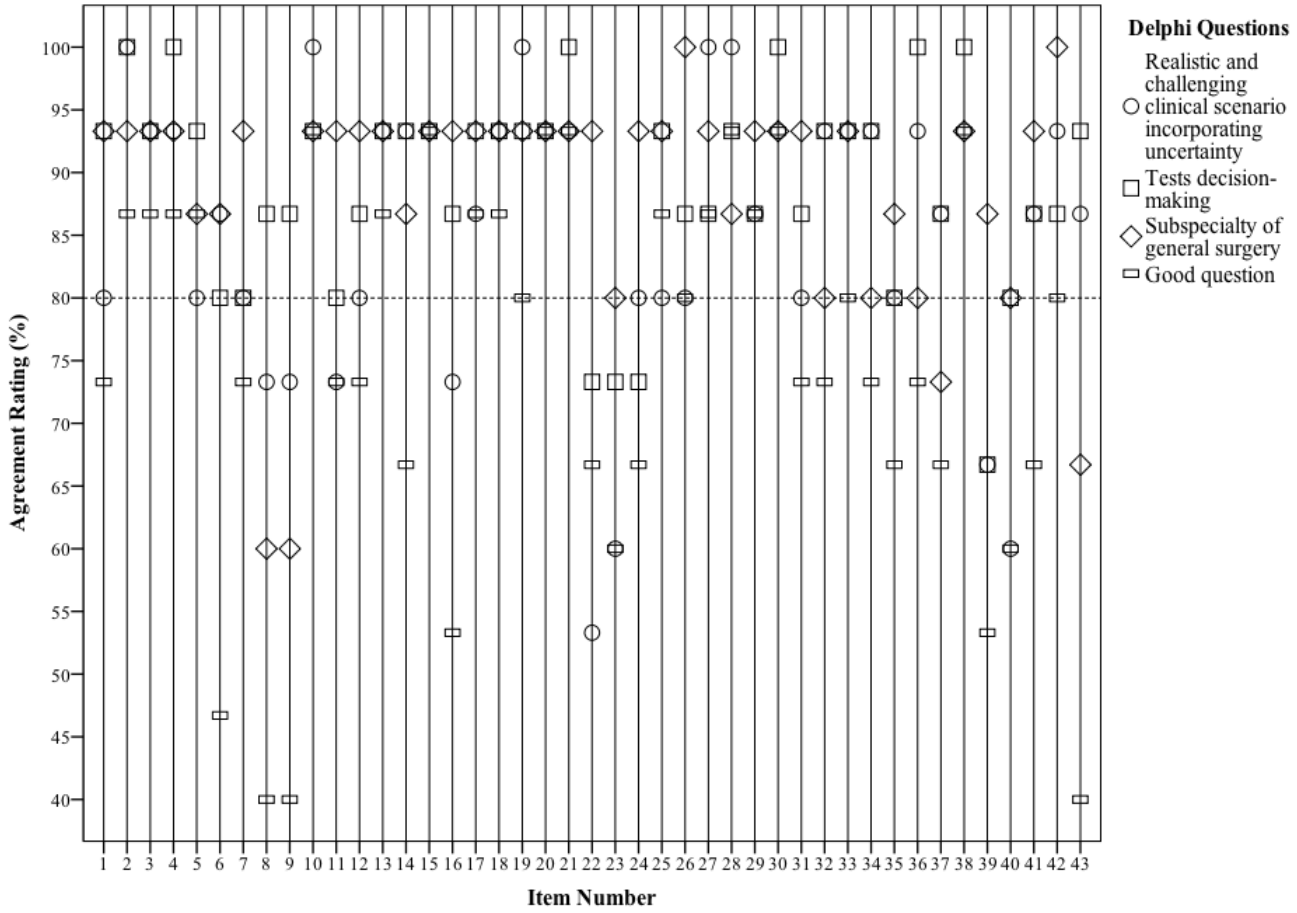

Figure 2. *Results of Expert Agreement for SCT Items in Delphi Round 1. Reference line at 80% depicts defined minimum agreement rating for consensus. Consensus must be achieved for all four Delphi questions for case to be included in final SCT*

Since only one round was necessary to achieve pre-defined consensus, all the written comments from Delphi question five (above) were reviewed and used to make changes only to unequivocal errors noted by one or more expert panellists, such as if there was a missed typographical error or an incorrect factual statement. Changes were not made based on comments that reflected a matter of opinion or otherwise changed the meaning of the case. All proposed changes were reviewed by another subject matter expert (Isabelle Raîche) prior to making the changes to confirm there would be no change to the overall meaning of the case or question with the proposed change.

Table of specifications.

The final table of specifications is presented in Table 6. All questions pertained to intraoperative management. No questions related to diagnosis or investigations. Many questions pertained to multiple areas of general surgery, and thus were mapped to multiple competencies. For example, a question about colon cancer was considered to encompass Colon and Malignancy competencies. The general surgery competencies that were not covered on the final SCT were Esophagus; Liver, Biliary Tract, & Pancreas; Breast; Adrenal Gland; Head & Neck; Pediatric Surgery; and Minimally Invasive Surgery

Table 6 *Table of Specifications for Final SCT (post-Delphi)*

Case & Question Number	Esophagus	Stomach & duodenum	Small intestine	Colon	Rectum & anus	Liver, biliary tract,	Spleen	Lymph nodes	Breast	Malignancy	Adrenal gland	Abdominal wall & hernia	Skin & soft tissue	Vascular	Head & neck	Trauma	Pediatric surgery	Minimally invasive surgery	Palliative care	Nutrition
1,1			X																	
1,2			X																	
1,3			X																	
1,4			X																	
2,5				X																
2,6				X																
2,7				X						X										
3,8				X						X										
3,9				X						X										
3,10				X						X										
4,11												X								
4,12												X								
4,13				X								X								
5,14				X						X										
5,15				X						X										
5,16				X						X										
6,17					X					X										
6,18					X					X										
6,19					X					X										
6,20					X					X										

Case & Question Number	Esophagus	Stomach & duodenum	Small intestine	Colon	Rectum & anus	Liver, biliary tract, pancreas	Spleen	Lymph nodes	Breast	Malignancy	Adrenal gland	Abdominal wall & hernia	Skin & soft tissue	Vascular	Head & neck	Trauma	Pediatric surgery	Minimally invasive surgery	Palliative care	Nutrition
7,21								X		X			X							
7,22								X		X			X							
8,23			X	X																
8,24			X	X																
8,25			X	X																
8,26			X	X																
9,27				X																
9,28				X																
9,29				X																
9,30				X																
10,31			X																	
10,32			X																	
10,33			X																	
10,34			X																	
11,35					X															
11,36					X															
11,37					X					X										
12,38								X				X								
12,39			X									X								
12,40												X								
12,41												X								
13,42				X																
13,43				X																
13,44				X						X										
13,45				X																
14,46		X												X						
14,47		X												X						
15,48			X																	X
15,49			X	X																
15,50			X																	
15,51			X																	
16,52			X											X						
16,53			X											X						
16,54			X											X						
17,55																X				
17,56																X				
17,57																X				
17,58																X				
18,59							X									X				
18,60							X									X				
18,61							X									X				

Case & Question Number	Esophagus	Stomach & duodenum	Small intestine	Colon	Rectum & anus	Liver, biliary tract, pancreas	Spleen	Lymph nodes	Breast	Malignancy	Adrenal gland	Abdominal wall & hernia	Skin & soft tissue	Vascular	Head & neck	Trauma	Pediatric surgery	Minimally invasive surgery	Palliative care	Nutrition
19,62				X												X				
19,63				X												X				
19,64		X														X				
19,65														X		X				
20,66												X								
20,67			X									X								
20,68			X									X								
21,69		X		X						X										
21,70		X						X		X									X	
21,71		X								X									X	
21,72		X								X										
Total	**0**	**7**	**22**	**26**	**7**	**0**	**3**	**4**	**0**	**19**	**0**	**10**	**2**	**6**	**0**	**11**	**0**	**0**	**2**	**1**

Table 7 demonstrates the general surgery competencies that were covered on the original vs. final SCT. Questions pertaining to the Esophagus; Liver, Biliary Tract, & Pancreas; Breast; and Head & Neck competencies were covered in the original SCT but did not achieve consensus during the Delphi and thus were excluded from the final SCT. Adrenal Gland, Pediatric Surgery, and Minimally Invasive Surgery were not covered on the original SCT and thus could not be included in the final SCT. Of note, minimally invasive refers to the operative technique that uses keyhole incisions and with the aid of a camera, as opposed to traditional open surgery which uses larger incisions such that surgeons can manipulate tissues with their hands (Townsend, Beauchamp, Evers, & Mattox, 2016). As such, the SCT is better described as a test of intraoperative management of open, adult general surgery to account for the lack of pediatric and minimally invasive surgery questions.

Table 7 *General Surgery Competencies Covered in Original and Final SCT Questions*

Competency	Original SCT	Final SCT
Esophagus	5	0
Stomach & duodenum	8	7
Small intestine	26	22
Colon	26	26
Rectum & anus	12	7
Liver, biliary tract, pancreas	21	0
Spleen	5	3
Lymph nodes	17	4
Breast	10	0
Malignancy	68	19
Adrenal gland	0	0
Abdominal wall & hernia	24	10
Skin & soft tissue	12	2
Vascular	9	6
Head and neck	19	0
Trauma	19	11
Pediatric surgery	0	0
Minimally invasive surgery	0	0
Palliative care	2	2
Nutrition	1	1

A range of themes emerged to explain why various competencies were excluded through the Delphi process. Comments written by surgeons for questions pertaining to the Esophagus (Original SCT Questions 142-144) and Head & Neck (Original SCT Questions 112-114, 119-134) competencies suggested they were excluded because the Delphi panel did not judge them to be relevant to the domain (general surgery). Comments included:

"Not sure if one would be expected to do a neck exploration as a general surgeon" – SD09

"It is a good question, but I would consider this now to be thoracic surgery, very few general surgeons were going to be involved with carcinoma of the EG junction." – SD12

"Probably not relevant to general surgery" – SD07

"Very specialized, most general surgeons don't do thyroid" – SD04

> *"…falls outside the realm of most general surgeons"* – SD01

For questions relating to the Liver, Biliary & Pancreas competency (Original SCT Questions 18-24 and 53-57) the Delphi panel's rationale for exclusion generally stemmed from the level of difficulty of the cases and questions presented. Comments from non-hepatobiliary (HPB) surgeons included:

> *"More appropriate for subspecialty level - HPB - general surgeon would not perform Whipple"* – SD09

> *"This is a detailed HPB question that is not suitable for a General surgery resident"* – SD05

> *"I think generally the Royal College has felt that decision making about Whipple procedures is not within the realm of surgical residents, the assumption is they will not be doing them without further training. I just find the question too hard for a general surgery resident, much more suitable for a hepatobiliary fellow"* – SD12

The sentiment of inappropriate level of difficulty was shared by hepatobiliary (HPB) surgeons, with added concerns as to the quality of the case and questions:

> *"It is a good question in principle, although it may be beyond the scope of a general surgery resident. The portal vein and mesenteric vein involvement needs to be qualified better. What does involvement mean? It is one thing to resect a small wedge of the PV if there is some tumor abutment. It is another thing entirely to find complete encasement of the PV with significant varicosities. The key qualifier is "reconstructible". Most HPB surgeons will proceed if they feel the PV or SMV is reconstructible."* – SD15

> *"…[questions] 20-24 controversial, and more HPB fellowship level"* – SD03

The Delphi panel's rationale for excluding questions related to the Breast competency (Original SCT Questions 1-4, 37-42) stemmed from an entirely different reason. For these questions, the Delphi panel generally felt as though the intraoperative findings presented in the question were unrealistic. In general, they explained that unexpected findings are extremely uncommon in breast findings due to the quality of the pre-operative imaging, and thus these scenarios with intraoperative findings not demonstrated on pre-operatively rendered the questions unrealistic. Comments from breast and non-breast surgeons both alluded to this rationale and included:

> *"I think the palpable lymph node is not a realistic option. This would have been discovered pre-op not intra-op therefore is misleading."* – SD05
>
> *"not very realistic mobile lump vs attached to fascia axillary lymph node should have been detected pre-op"* – SD14
>
> *"…I would only comment that in modern breast surgery, finding something at operation that was not anticipated from the preoperative imaging is virtually impossible nowadays. changing the procedure based on intraoperative findings is very uncommon now."* – SD12

Summary of Results for Test Content Validity Evidence

In summary, this chapter described how a Delphi and table of specifications were used to establish test content validity evidence. A Delphi panel of 15 local surgeons voted to be the top decision-makers with the Division of General Surgery was formed and answered questions relating to items from an existing SCT that had been previously administered to general surgery residents nationally (Nouh et al., 2012) with the goal of achieving 80% consensus for 20 cases. Consensus was achieved for 21 cases after the first round of the Delphi. The post-Delphi SCT served as the final SCT for use in this study, and questions were mapped to general surgery competencies as defined by the RCPSC on a table of specifications.

The table of specifications demonstrated overall good coverage of the intraoperative management of general surgery, with some notable missing competencies. No questions relating to adrenal glands, pediatric surgery, or minimally invasive questions were present on the original or final SCT. Questions relating to Esophagus and Head & Neck competencies were present on the original SCT but excluded from the final SCT due to perception of these questions being outside the scope of general surgery as per the Delphi expert panel. Questions relating to the Liver, Biliary, & Pancreas competency were excluded for being too difficult, and therefore more

appropriate for the HPB fellowship level. Questions relating to the Breast competency were excluded for being unrealistic given the presentation of unexpected intraoperative findings in the context of excellent pre-operative imaging in modern day breast surgery. Although a second Delphi round could have been employed to obtain consensus on cases encompassing all competencies in the table of specifications, in addition to ensuring overall good coverage of general surgery, the intended purpose of mapping the competencies was to identify and explain why some competencies may not be included. Furthermore, obtaining consensus on excluded cases would have involved writing new cases which would not have undergone the same psychometric testing as the original cases from the SCT used in the national validation study (Nouh et al., 2012). The final SCT was then used to establish response process validity evidence, which is presented in the next chapter.

Chapter 3: Response Process Validity Evidence

Overview

Response process validity evidence is the second of the five sources of validity evidence presented by the *Standards for Educational and Psychological Testing* (AERA et al., 2014). Along with content validity, response process validity evidence is typically gathered during the development of any assessment tool (AERA et al., 2014), prior to pilot testing. Response process evidence is the extent to which the thoughts of test-takers demonstrates their understanding of the construct as it is defined (Padilla & Benitez, 2014). Obtaining response process evidence can be done by one of two methodological categories: direct assessment of cognitive processes by questioning test-takers, or indirect indicators of performance such as eye movement or response

time (Padilla & Benitez, 2014). The study in this thesis sought to obtain response process evidence by conducting cognitive interviews to examine the issues with respect to the cognitive processes participants use to answer SCT questions (AERA et al., 2014).

Cognitive interviewing explicitly focuses on the cognitive processes that respondents use to answer questions. The most commonly used and general model is attributed to Tourangeau (1984) and consists of the following processes: comprehension of the question, retrieval of relevant information from memory, decision-making processes, and response processes. Since the SCT items are non-trivial, the question-answering process can be complex and involve a number of conscious and/or automatic cognitive steps. (Willis, 1999) As such, multiple cognitive codes may be assigned to each item.

Methods: Response Process Validity Evidence

Sampling and participants.

After development of the final SCT to be used for this study, the test was administered to a purposeful sample of 14 staff general surgeons, all of whom had participated as members of the Delphi panel. Of note, the fifteenth Delphi panellist was unable to continue to participate as an SCT expert panellist due to time constraints. The 14 surgeons forming the expert SCT panel were identified to be the top clinical decision-makers in an informal survey administered to all staff and resident general surgeons. The size of the expert panel is congruent with SCT recommendations in previously published literature (Fournier et al., 2008).

The final SCT was also administered to a convenience sample of 29 of the 35 resident surgeons within the Division of General Surgery at the University of Ottawa. Each resident and staff participant was assigned a random number to ensure their anonymity. In addition to gender,

post-graduate year (PGY) and clinical subspecialty were the only other demographics collected for resident and staff participants, respectively. Of note, response time could not be reliably collected as test-takers could take the SCT in a single sitting, leave it open while doing other work, or close and re-open the test as per their availability.

After completion of the SCT, ten resident surgeons and five staff surgeons participated in cognitive interviews. A diagram of the study design is presented in Figure 3.

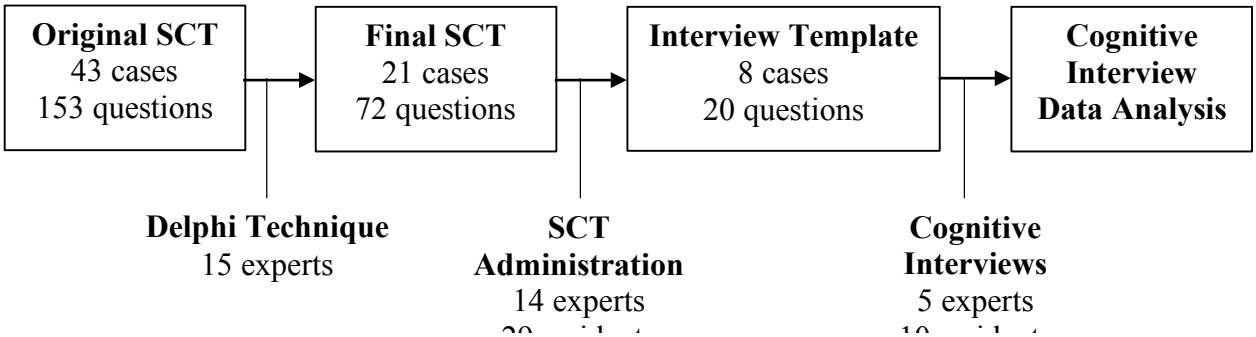

Figure 3. *Study Design*

The number of interviews that could be performed was limited by feasibility and time. Thus, the sample size of ten resident surgeons and five staff surgeons was ultimately determined as a purposeful sample of surgeons within our division to find outliers that could confirm previous findings or enrich the data with their points of view (Marshall, 1996). Since the expert panel taking the final SCT was the same composition as the previously described Delphi panel (minus one surgeon), the clinical subspecialties of the expert panel are the same as those presented in Table 4, minus one breast surgeon. From the expert panel, the surgeons participating in the cognitive interviews consisted of three males and two females. These five surgeons represented five different clinical subspecialties and practice at both campuses of The Ottawa Hospital. Their practice experience ranged from two years to over thirty years.

The resident surgeons taking the final SCT consisted of six PGY1, seven PGY2, six PGY3, five PGY4, and five PGY5 residents. Of these residents, 17 were male and 12 were female. From the residents that took the final SCT, ten participated in cognitive interviews. These ten resident surgeons consisted of one female and one male resident from each PGY level, for a total of five females and five males, with two residents at each PGY level (four junior residents, six senior residents).

Cognitive interviews.

The cognitive interview is a qualitative method that uses a semi-structured interview protocol to guide the interview and access the cognitive processes of test-takers (Padilla & Benitez, 2014). A retrospective verbal probing technique was used. Verbal probing is a technique in which the interviewer probes into the basis for the test-takers' responses and was used as opposed to the think-aloud technique because it does not require participant training, it allows the interviewer to have greater control over the interview, and it shifts the burden to the interviewer instead of the participant (Willis, 1999). To minimize the potential for bias, non-leading probes were carefully selected and scripted using unbiased phrasing for each item (Appendix 2. Cognitive Interview Guide). Probes were selected in accordance with cognitive interviewing guidelines by Willis (1999), ensuring that all four stages of Tourangeau's cognitive model (1984) were represented (Table 1). The scripted probes were then reviewed by a cognitive psychologist (Katherine Moreau). However, there was some latitude in the execution of probes to allow for spontaneous and emergent probing (Willis, 2015). The cognitive interviews were conducted retrospectively mainly because of the mental effort by participants required to answer the items. Relatedly, the significant mental effort required for a large number of items precluded the inclusion of all items in the cognitive interview, and thus which items to include in the

cognitive interview were carefully selected based on the graphically depicted results after administration of the final SCT. Items for inclusion were selected to include a variety of response patterns between resident and expert (staff) surgeon test-takers (

Table 8). As shown in

Table 8, results were reviewed and selected based on agreement within the staff group, agreement within the resident group, and agreement between the staff and resident groups. Agreement was defined as questions for which most respondents selected some degree of indicated (+1, +2), some degree of contraindicated (-1, -2), or neither more or less indicated (0), but were not divided between these categories (i.e. -1, +1). Of note, there were no results demonstrating a lack of agreement among staff, but agreement among the residents, and thus no questions demonstrating this pattern were available for selection. The final cognitive interview guide was used for 20 pre-determined questions stemming from eight cases on the SCT. The cognitive interview guide spanned eight of the 13 general surgery competencies included in the final SCT's table of specifications.

Table 8 *Response patterns of questions selected for cognitive interview guide*

Staff agree	Residents agree	Staff agree with residents	Common responses	Number of Questions
Yes	Yes	Yes	+1, +2	3
Yes	Yes	Yes	-1, -2	3
Yes	Yes	Yes	0	1
Yes	Yes	Yes	Staff: 0, -1, -2 Residents: -1, -2	1
Yes	Yes	Yes	Staff: 0, -1 Residents: 0, -1, -2	1
Yes	Yes	No	Staff: -1, -2 Residents: +1, +2	1
Yes	Yes	No	Staff: 0 Residents: -1, -2	1
Yes	No	--	Staff: +1, +2 Residents: all	1
Yes	No	--	Staff: 0 Residents: +1, -1	1
Yes	No	--	Staff: +1, +2 Residents: -1, 0, +1	1
No	Yes	--	--	0
No	No	Yes	All	1
No	No	Yes	Staff: all Residents: -1, +1	2
No	No	Yes	Staff: -1, 0, +1 Residents: all	1
No	No	Yes	-1, +1	1
No	No	Yes	-1, 0, +1	2

It should also be noted that the option presented in the probe "what would have made you choose ___ option?" was pre-selected to be either the opposing modal response on the final SCT, or the more extreme or more moderate response of the same directionality, depending on the question. For example, if a respondent arrived at answer of -1 (contraindicated), and a significant number of responses on the final SCT were -1 and +1 (indicated), it might be asked "what would have made you choose +1?" to probe the opposing viewpoint. Alternatively, if a respondent arrived at an answer of -2 (strongly contraindicated), it might be asked "what would have made you choose -1" to probe the reasoning the contraindication was thought to be strong. In addition, several modifications were made to the interview guide during the study. After the first two

interviews, it was noted that interviewees were simply re-stating the question when answering the first probe “in your own words, what is this question asking you?”, as such this probe was re-worded to “what is this item testing?”. Second, it was noted that interviewees tended to jump straight into the third question “how did you arrive at the answer you chose?” and thus this question was asked first in subsequent interviews.

Data collection.

I conducted all the cognitive interviews. I have a pre-existing workplace relationship with all the participants, which results in both advantages and biases with respect to data collection. The nature and influence of this relationship is discussed in greater detail in the *Interview data collection and analysis team*. section below. Each cognitive interview was scheduled for 60 minutes and ranged from 50 to 115 minutes in duration. The length of interview time generally reflected participants’ depth of explanation and note of other perspectives. The entire interview guide was applied to all 20 questions in 14 of the 15 interviews. In the one interview that did not include all questions, the interviewee described fatigue with the cognitive load of the probes, and requested to focus the remaining questions to complete the interview within the 60 minutes 12 were conducted in person at the location of the participant’s choice, and three were conducted via Skype or Facetime due to the participant’s location during periods of availability. All interviews were audio-recorded and I typed the interviewee’s responses under each question during the interview, including notes on non-verbal behaviours.

Data analysis.

Data was analyzed using an approach proposed by Willson and Miller (2014), which is suggested to be particularly useful in conducting validation studies aimed at obtaining response process validity evidence (Padilla & Benitez, 2014). First, notes generated during the interview on a question by-question basis (Willis, 1999) were used as the basic unit of analysis. These findings were analyzed using *a priori* "top-down" deductive cognitive codes (Willis, 2015). Deductive cognitive coding involves the application of coding categories prior to data analysis (Willis, 2015). For each item, one or more cognitive codes were assigned to code any problem faced by the test-taker as a problem of one or more of Tourangeau's cognitive model: comprehension, recall, decision-making, or response matching (1984). It should be noted that unlike problems identified with respect to comprehension, recall, and response matching, findings coded as decision-making do not represent problems in the SCT context. In fact, evidence of a decision-making process ties into the construct validity of the SCT by demonstrating that respondents are using a decision-making process to answer SCT questions as opposed to relying strictly on recall of knowledge.

Cognitive codes were assigned independently by me and one RA (Lindsay Cowley). Inter-coder reliability was not quantified as discrepancies were thought to be an asset, by demonstrating multiple lenses through which the data could be interpreted. However, no major discrepancies (i.e. altering the modal cognitive code for an item or for a test-taker) occurred, and when minor discrepancies were noted (i.e. for a single test-taker on a single question) the cognitive codes assigned by both coders were used. Data analysis was an iterative and constant comparative process involving regular meetings with my clinical research supervisor (Isabelle Raîche) to discuss summarized findings and emerging themes. Next, comparison across

respondents was used to identify which cognitive code(s) each SCT item encompasses. Then, the cognitive processes of groups, namely junior residents (PGY1 & 2), senior residents (PGY3, 4, & 5) and staff surgeons (expert panel), were compared looking for similarities and differences. The cognitive processes of those whose responses were concordant with the expert panel were also compared to those whose responses were discordant with the expert panel to elicit any differences in their cognitive processes. No software was used during data analysis.

Interview data collection and analysis team.

As the principal investigator conducting the cognitive interviews and analyzing the data, I must acknowledge my personal biases which may influence the results of this study. As a PGY3 general surgery resident, my position is that of an insider-outsider relative to the study participants. My training provides me with a lens that is likely similar to that of the participants, particularly with respect to the resident participants. As per the social hierarchy that exists within the culture of surgical training, all the staff, PGY5, and PGY4 participants were in a position of power over the researcher. The PGY3 participants were on a similar level within the social hierarchy, while the researcher was in a position of power over the PGY1 and PGY2 participants within the hierarchy. The advantage of this pre-existing relationship is the easily established rapport because of a pre-existing workplace relationship, as well as the ability to use a common language and draw on a common experience base (McDermid, Peters, Jackson, & Daly, 2014). However, a pre-existing workplace relationship may also pose risks in terms of the quality of the data collected and analyzed (McDermid et al., 2014). Steps taken to ensure quality of the data are described in the next section, titled *Trustworthiness*.

Because of the potential influence of the pre-existing relationship on the data analysis, data was independently coded by an RA with no surgical training or relationship with the participants (Lindsay Cowley), and as such may view the data with a completely different lens. The RA who conducted the secondary analysis holds a Master's Degree in Applied Linguistics, which affords her a unique and complementary perspective.

Trustworthiness.

Lincoln and Guba's four criteria for trustworthiness was used to ensure quality data collection and analysis: credibility, transferability, dependability, and confirmability (Lincoln & Guba, 1985). Credibility refers to the confidence in the truth of the results. Credibility was achieved by triangulating the participants in terms of level of training/years in practice, subspecialty of practice, campus, and gender as well as using two independent data analysts with different perspectives. Transferability is the applicability of the findings to other similar contexts. Interviews were recorded and descriptions of research settings and participants were provided to allow readers to evaluate the extent to which the studied context is transferable to their own. Dependability is defined as the consistency and repeatability of the findings. Dependability was promoted by allowing the second analyst and a second subject matter expert to evaluate and challenge the findings reported and conclusions drawn. Finally, confirmability reflects the neutrality of the findings, or the absence of bias. Confirmability was achieved by triangulating participants and analysts as described above, selecting unbiased probes and top-down coding (Willis, 2015), documenting analytic decisions, and implementing a reflexive process. Reflexivity included introspection of personal biases, reviewing and refining interview technique on an ongoing basis, discussing multiple perspectives with the research team (Finlay, 2002). To

minimize the impact of the inevitable bias due to the pre-existing workplace relationship, participants were assured that their responses would remain confidential. In addition, I clearly explained that the purpose of the interview is to explore their thought process during the SCT, not to assess their CDM skills. Participants were encouraged to say they didn't know something rather than guess, and the value of each of their unique positions as juniors, seniors, and staff surgeons was reiterated.

Results: Response Process Validity Evidence

Overview.

In exploring the cognitive processes of test-takers, it became apparent that test-takers relied on scripts formed through past experience to make decisions. In addition, issues pertaining to all four stages of Tourangeau's cognitive model were observed within various SCT items. In the context of response process validity evidence, the dual-process theory and the cognitive model are used as frameworks within which to identify response process validity evidence for various SCT items. As described in AERA's standards and testing issues for evidence based on response processes, Standard 1.8, since the rationale for the SCT's use and score interpretation depends on premise about the cognitive processes used by test-takers, then theoretical or empirical evidence supporting those premises should be provided (Padilla & Benitez, 2014). After presenting evidence for the dual-process theory of CDM, this chapter describes the results of the cognitive interviews in sections divided by each of the four mentioned cognitive processes. Each issue identified is accompanied by possible and researcher-proposed solutions that serve as next steps and each warrant future investigation. Saturation was deemed to be

reached, as no new themes emerged from the data after completion of the interviews (Padilla & Benitez, 2014).

Dual-process theory.

Dual process theory describes System 1 and System 2 thinking which represent non-analytic and analytic based reasoning, respectively. Scripts are an example of a System 1 non-analytic resource, where the decision-maker uses experience to activate a relevant script, thus promoting System 1 thinking. However, decisions are seldom made exclusively by System 1 or System 2 thinking; although one may predominate, the two systems interconnected. Cognitive interviews demonstrated that when test-takers had relevant experience to draw on, that was the predominant factor in their decision-making process. For example, in a question asking about the management of a lower gastrointestinal bleed (LGIB) from an unidentified source, a senior resident described a similar previous case in their decision:

> *"My understanding is…you would still proceed with subtotal colectomy…I'm leaning on previous experience where we did a subtotal and there was blood from matured stoma, and it still worked to control the LGIB!"* – SSRES04

This senior resident described the question as easy and the extent of the explanation of their response choice was limited to their previous experience. This demonstrates predominantly System 1 thinking, with a reliance on a formed script for managing LGIB of an unidentified source. On the other hand, another resident of the same level of training, but who had never seen a case of LGIB of an unidentified source responded to the same question by analyzing the principles:

> *"Okay. So. This is indicative to me of a more proximal site of bleeding than the colon, makes me think a subtotal is unlikely to be effective. Is there a possibility its backwash bleeding with a colonic site? So I'm debating between -2 and -1. In particular, a subtotal alone is not an appropriate operation. [This is] difficult. In my mind, I don't have a clear answer of what I would do. I don't have any experience of operative management of LGIB, so I have no clinical experience to drawn on to answer this."* – SSRES01

Despite the same level of training, because of no previous experience with the case, the latter senior resident relied on analysis of surgical principles in determining his decision. He did not find the question easy as did the former senior resident, and specifically cited his lack of experience as the difficulty in making the decision.

This finding was not unique to this item. In general, senior residents and staff tended to rely on experience, when available, to make their decisions. Junior residents, who have less experience, tended to analyze their knowledge to arrive at a decision. These response process findings support the dual process theory of decision-making. Specifically, the findings support the predominance of script-based decision-making (System 1) when the test-taker had relevant experience to draw on versus predominantly analytic decision-making (System 2) when the test-taker did not.

Comprehension.

The comprehension process of Tourangeau's cognitive model pertains to both the meaning of specific terms and the overall question intent (Willis, 1999). Specifically, it ascertains if specific words and phrases have the same meaning to participants as they do to those making the test, as well as whether participants' understanding of what the question is asking is congruent with test-makers' intention. The comprehension issues identified with the SCT mainly stemmed around three themes: varying interpretation of the term 'no intervention', the desire for additional detail in the item, disagreement with the initial planned management.

There was a single item that demonstrated variable understanding of the meaning of a specific term across multiple participants. There was no other indication of misunderstanding or

difficulty understanding specific terms. The item is shown below (Figure 4), with the specific term in question being "no intervention" in the proposed management plan:

Scenario 3)

You have been called into the Gynecology operating room for an intra-operative consultation. The patient is undergoing a scheduled open hysterectomy for menorrhagia. During the dissection, the gynecologist has found a mass in the sigmoid colon. The remainder of the pelvis and what can be visualized through the lower transverse incision is normal.

If you were planning…	and in the OR found…	the planned management is:				
9) No intervention at this time with a work up post-operatively	there is a small amount of stool leaking from an colotomy made during the dissection, located 2 cm proximal to the mass	-2	-1	0	+1	+2

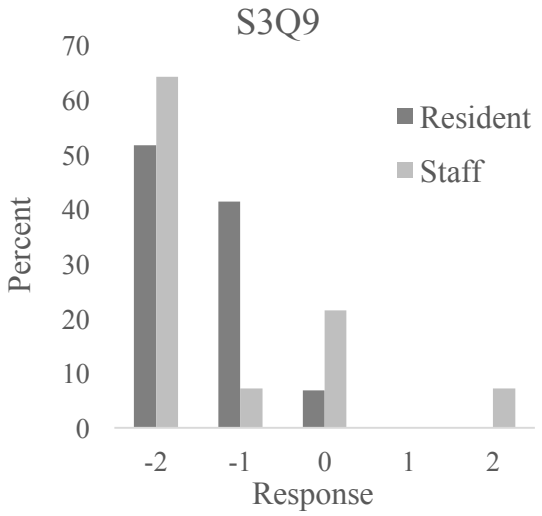

Figure 4. *Scenario 3, Question 9 with graph demonstrating resident and staff responses*

When participants were probed as to how they arrived at the answer they chose, it became clear that participants understood "no intervention" to either mean absolutely nothing, or they thought repairing the colotomy was so obviously indicated that the question must be referring to no intervention on the mass (i.e. no resection) after repairing the colotomy:

> *"if no intervention means not even closing the hole that's a major problem—does this mean I'm not dealing with the mass or not even getting source control of the colotomy made?"* – SRRES02

> *"To me no intervention means I'm going to fix the colon, obviously, but I'm not going to resect the colon."* – SRRES03

> *"I wouldn't want to leave [the colotomy] open. I don't like the question. I want to intervene on the hole, not on the mass."* – JRRES01

> *"No intervention to me means you've repaired the colotomy…it's asking me if I'm going to do a resection. But I'm not going to do a resection, I'm going to primarily repair the colotomy. Which is not no intervention, it's repairing the colon…I think they're asking are you going to do a resection?"* – EPSS01

> *"No intervention means you're getting out [without repairing the colon]. If there's stool coming out and you say see you later, you're getting sued. Does that make sense to anybody?"* – SRSS01

> *"There's a lot of double negatives here [when you ask if I don't want to do nothing]…I seem to have trouble figuring out what you were asking"* – SRRES05

The comprehension issue for this item stemmed from there being two possible management plans: management of the colotomy and/or management of the mass. The proposed plan of "no intervention" was interpreted by some participants to be with respect to the colotomy and mass, and interpreted by other participants to be with respect to just the mass assuming repair of the colotomy was given. Some participants could see both interpretations, which generated confusion as to the intent of the question. One possible solution for the variable interpretation of "no intervention" is to change the planned management to an operative procedure.

The second issue with respect to comprehension was elicited for multiple items, in response to the probe "was there any information you didn't know that you felt you needed to answer the question?". In response, participants frequently requested more detail about the question. An example of this is shown in the question below (Figure 5):

Scenario 10)
You are operating electively on a 39 year old male with Crohn's disease of the small bowel and obstructive symptoms.

If you were planning…	and in the OR found…	the planned management is:				
31) To do a small bowel resection	3 strictures all in close approximation to each other	-2	-1	0	+1	+2

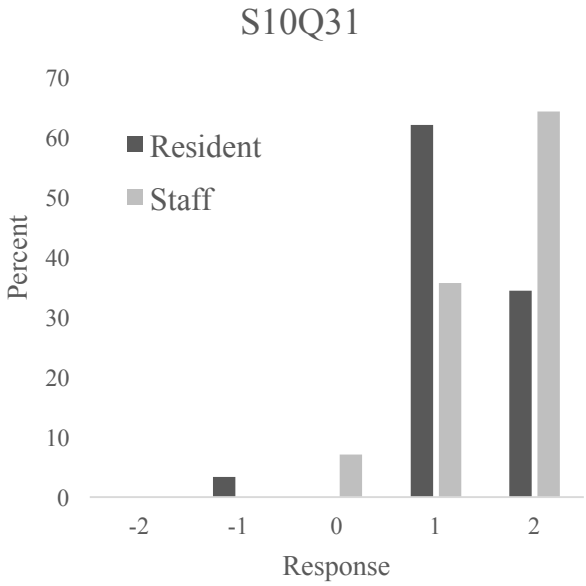

Figure 5. *Scenario 10, Question 31 with graph demonstrating resident and staff responses*

Here participants frequently wanted to know how close the description of "close approximation" was and indicated that this would change their indication to do a small bowel resection:

> *"…what does 'close approximation' mean? ...I would probably say +1, the fact that there's 3 in close approximation makes me think resection of the entire area is physiologically tolerated. But [it would be] nice to have a bit more information…if it said total involvement of 10cm then it would be +2 and easy."* – EPSS02

However, despite frequently seeking more information, there was also an understanding that the desire for more information exists may be inherent to the ambiguity of the SCT:

> *"With any of these types of tests there's always information you don't have…It's very hard to differentiate that with a one-liner. I think in my head 'how would I write this better?' Because if I explain all [the information I want to know], I've given the answer away…you can only provide so much. I get that."* – EPSS01

> *"No information about comorbidities implies he's healthy because they didn't tell you he isn't. With [these questions] you have to assume, and that's reasonable. So there's nothing more I need."* – SRSS01

Such responses suggesting certain assumptions are inherently necessary, as well as the observation that the desire for more information seemed to be more specific to participants and

not consistent across all participants for any given item, suggests that no revision to provide additional details are necessary. However, given that this was a source of frustration for some participants, including a description that the SCT purposefully includes ambiguity may be valuable to frustrated participants.

The final comprehension theme occurred for two items where several participants did not agree with the proposed management plan. For example, in an item proposing an exploratory laparotomy 24 hours after presentation with a small bowel obstruction, many participants did not think the exploratory laparotomy should be done so soon in the absence of explicitly stated peritonitis. Other participants assumed the patient must be peritonitic to warrant the planned management. For another item proposing an elective hernia repair in an obese smoker, many participants described the intraoperative finding to be irrelevant because they would not have operated on this patient at all:

> *"This is the [question] I'm the least happy about in terms of understanding why I'm there…the [plan] started differently than I would treat them, so I'm confused as to why I'm there in the first place…I really rely on the stem to give me a clear indication of my plan, and when [my mental plan and the test's proposed plan] don't line up it adds a level of challenge because I'm also like 'wait why would I have done that in the first place? And I found something I wasn't expecting, but why am I there in the first place?' – there are two layers to the question."* - SRRES05

For these items, participants who did not agree with the initial planned management could not comprehend how to respond to the question since it was not the intraoperative finding that made them disagree, as was thought to be intent of the question, but rather the initial planned management. A proposed solution to this issue is to ensure consensus among the expert panel with respect to the planned management prior to test administration.

In summary, cases and questions were generally well written with clear terminology that was accurately comprehended by participants. The exception to this was the term "no

intervention", where participants had varying interpretations as to the meaning. Relevant additional information was frequently sought by participants to aid in their decision-making process, but they generally understood that some level of ambiguity is inherently necessary to make the question difficult to answer. Finally, participants did not comprehend how to answer questions in which they did not agree with the initial planned management. They thought the new intraoperative finding should guide their response, not the initial planned management.

Recall.

The second stage of Tourangeau's cognitive model is recall, which pertains to the retrieval of relevant information from memory and the strategies used to retrieve information (Willis, 1999). Recall of information assumes that it was learned in the first place, which is fundamentally variable given the diverse levels of training included in this study. Accordingly, responses were more frequently coded as "recall" issues for junior level trainees, likely because the relevant knowledge had not yet been acquired. The expert group of staff surgeons therefore serve as an ideal group to discern issues with retrieving the relevant information from memory, because it is reasonable to assume that they have learned the relevant information at some point.

Issues with items which were coded as predominantly recall for staff surgeons were those relating to very subspecialized areas of general surgery, such as melanoma. At The Ottawa Hospital, there is only one general surgeon that sees melanoma cases in their practice out of the 25 general surgeons on staff. The verbal probe "did you find this question difficult or easy?" was used to elucidate the recall cognitive stage being the predominant issue with respect to items pertaining to melanoma, among other very subspecialized items. Staff surgeons unanimously described these items as difficult:

> *"Oh this is not my alley…I haven't looked at this stuff in ages…I'm a little bit confused, maybe because I don't deal with melanoma very often…I don't have solid ground because I don't deal with this to be honest. Maybe [this isn't difficult] for someone who deals with melanoma, but it's difficult for me."*- SRSS01

> *"I don't have a great knowledge of melanoma management. But I would think, I believe you can have separate draining basins, and if it's central, it seems reasonable there are two draining basins. Like in anal cancer, so I'm extrapolating. [This is] difficult, mainly because I don't treat melanoma—ever."* – EPSS03

Staff surgeons use an estimation strategy to attempt to answer questions for subspecialized areas for which they lack knowledge. By applying the knowledge of their own area of practice, they extrapolate basic oncologic principles to attempt to answer a question about another area of practice, as opposed to recalling the directly relevant knowledge. As with the above item pertaining to melanoma, the finding of recall issues with respect to subspecialized items was also noted for questions pertaining to stricturoplasties, which are generally only seen and performed by colorectal surgeons.

Items pertaining to truly general surgery that would be encountered in an acute setting by any subspecialty surgeon, for example when on call, did not demonstrate any recall problems. The identified recall problems with subspecialized items suggest revisions where either subspecialty items are eliminated, or the expert panel modified to allow subspecialty-specific scoring.

Decision-making.

The decision-making stage of Tourangeau's cognitive model consists of two aspects: motivation and sensitivity/social desirability. Motivation pertains to the mental effort the participant devotes to answering the question accurately and thoughtfully. Sensitivity and social

desirability refer to the participant's desire to answer in a way that makes them look "better". (Willis, 1999)

The majority of responses for most items were coded as "decision-making". Simple observation of participants made it clear that most items were difficult to answer. Participants took a long time to respond to each item, and tended to be unsure about many items, often weighing multiple perspectives and considerations before arriving at a final response. The significant cognitive load of the decision-making process was also observed directly through statements made by participants:

> *"[It's] difficult in that it's intraoperative decision-making. There are two clearly okay answers, one is probably more okay, but it's more of a grey zone."* – EPSS01

> *"the question wouldn't necessarily get easier with more knowledge, [it's] just a decision-making thing."* – JRRES01

> *"I can see how a trainee would find it challenging…I don't think [a different answer is] a failure of knowledge or a critical knowledge gap because I acknowledge there is controversy in this area."* – EPSS02

Unlike items coded as problems with respect to comprehension or recall, items coded as "decision-making" do not represent a problem at all since the intended purpose is to gather validity evidence that a decision-making step occurs. Of note, the motivation part of decision-making serves as validity evidence, but the social desirability part does not. Social desirability is not indicative of validity because ultimately it is important to determine the extent to which test-takers' responses are concordant with expert responses, regardless of why.

However, participants' decision-making process was influenced by social desirability to varying extents. Surgeons must strive to achieve a balance between a lack of confidence with overconfidence; they must demonstrate both sufficient confidence to make decisions that come with inherent risk and the humility to accept responsibility for their mistakes (Angelos, 2017). As

trainees learn to negotiate the confidence balance, in any given situation they may strive to exert confidence if they perceive a lack of confidence to be socially undesirable, or alternatively may strive to demonstrate humility if they perceive overconfidence to be socially undesirable. This study demonstrated that trainee participants more so desire humility and did not want to appear overconfident. Trainees would often justify their answer when responding to the probe "why did you pick +/- 1 instead of +/- 2" by citing their desire to appear moderate and a lack of confidence as reasons to not pick +/-2.

> *"On this type of exam I'd be more prone to choosing a more moderate number to assure myself I hadn't made a very bold statement"* – SRRES01

> *"To some degree, how I am using the scale is really just a representation of my confidence"* – SRRES04

The minority of trainee participants took the opposite approach; they strived to purposefully exert their confidence, instead perceiving unsureness as socially undesirable.

> *"This is someone with Crohn's…so it's completely inappropriate…no, [nothing would make me pick] -1…no, [nothing would make me pick] +1. [It's] easy."* – SRRES02

> *"[It] is unreasonable because it's not the best outcome…I am confident I don't want to resect en bloc…[that's] easy."* – SRRES02

Although it is difficult to illustrate that these responses are purposefully confident due to social desirability as opposed to due to knowledge, it becomes more apparent when the above responses are compared with the responses of experts who describe the same items as difficult due to multiple competing considerations. When the impact of social desirability on the decision-making process was observed, it was pervasive across most items for that participant and thus seemed to be more of a reflection of the participant. As such, the impact of social desirability was found to vary with individual participants, not with specific items. Participants demonstrating such confidence did so no matter how difficult experts reported the case to be.

Although the influence of social desirability was not observed in every trainee participant, it was not observed to influence the decision-making process of any staff participants. This finding could be considered an issue with respect to response process as it ultimately determines how participants choose their response. However, it was coded as decision-making because the underlying reason for the internally generated response is influenced by the effect of social desirability on the participant's decision-making process. How to alleviate the impact of social desirability on the decision-making process of participants remains unknown as it may be more overt in comparison to other written tests currently used in medical education.

Response matching.

The final category of Tourangeau's cognitive model is the response matching process, whereby participants match their internally generated answers with the response categories provided by the question (Willis, 1999). This category generated the most number of problems identified of all four categories, mostly likely due to the novel Likert-type scale utilized by the SCT. Five major issues were coded as response matching: the meaning of the 0 response, what response corresponds to agreement with the planned management, rationale for picking +/-1 versus +/- 2, what response indicates the desire to do the planned management plus an additional procedure, and the influence of time on response selection.

The SCT uses a Likert-type scale with response options of -2, -1, 0, +1, and +2. Each of these numbers is accompanied by a description of its meaning: strongly contraindicated, contraindicated, neither more or less indicated, indicated, and strongly indicated, respectively. Despite these consistent descriptors, the probe "how did you arrive at the answer you chose?"

demonstrated variable interpretation of the reasons for which the response 0 (neither more or less indicated) was selected.

> *"0 means I need more information, I just don't know."* – EPSS02
>
> *"0 represents I would just continue with the plan, but also [it represents that] I just don't know. With knowledge, I'd be able to pick a more definitive answer. In general [for all items], if I had knowledge, I would not pick 0 because I'd favour one way or the other."* – SRRES03
>
> *"So it warrants repair but I'm not keen on it [because of patient factors] so that balances out to 0 in my mind."* – SRRES03
>
> *"So I think I'd say 0 because you could argue for or against this option. 0 is because I'm kind of ambivalent, you could argue both."* – SRRES02
>
> *"I wouldn't pick 0. I don't like the 0."* – SRRES06
>
> *"0 means neither indicated or contraindicated. I don't understand how you stand on the fence with something surgical, you either do it or don't do it."* – SRSS01

The intent of the 0 response option is to indicate that the new finding does not influence the test-taker's decision to pursue the planned management (i.e. pursuing the planned management is neither more or less indicated given the new finding). While misinterpretation of its meaning could be considered a comprehension issue, the use of 0 was coded as a response process issue because participants modified its meaning to suit their internally generated response. As demonstrated, participants used the 0 response option when they could not internally generate a response due to the desire for more detail in the question (comprehension issue) or a lack of knowledge (recall issue). Participants also selected 0 when they determined the pros and cons to be of equal weight, or when they decided there was clinical equipoise for the item. Finally, other participants described disliking the 0 response option, or finding it unsuitable because it represented no decision. Possible solutions include clarifying the intent of the 0 response option in the SCT instructions, or omitting the 0 response option entirely.

The second response matching issue identified stems from which response option (0, +1, +2) corresponds to participants' internally generated response of agreeing with the planned management, such as in the item below (Figure 6):

Scenario 20)

A 50 year old obese male smoker presents with an incisional hernia and pain at the hernia site while working.

If you were planning…	and in the OR found…	the planned management is:				
66) hernia repair with mesh	simple 1.5 cm fascial defect	-2	-1	0	+1	+2

Figure 6. *Scenario 20, Question 66 with graph demonstrating resident and staff responses*

This issue was initially observed because participants would explain the same rationale for their answer choice, but would select different numbers:

> *"The defect [size] itself doesn't change my plan so I say 0, it's the patient's clinical picture that makes me want to use a mesh…his clinical picture really plays into my decision-making."* – JRRES03

> *"Ya I think that's fine. +2. It's a small defect, mesh is a good idea in an obese male."* – JRRES02

> *"I would choose +2 because there's a simple fascial defect >1cm in size. It's not that cut and dry, but 1.5cm in an obese smoker to me means mesh."* – EPSS02

> *"because he's that high risk I would still put a mesh because of him being obese and a smoker…0 is [the fact that it's a 1.5cm defect] means nothing."* – EPSS02

> *"Well, probably +1. He's a smoker, he has a higher chance, he's obese, higher chance of recurrence. People with mesh repair have lower recurrence rate. Because of patient factors."* – EPSS03

Participants' rationale for this question cited patient factors (obese, male, smoker) to be the reason for wanting to proceed with mesh, as opposed to the size of the hernia defect. The finding that different response options were indicated despite the same internally generated response was pervasive across levels of training, as is demonstrated among junior residents as well as staff surgeons. Unlike example of issues with respect to knowledge (recall) or social desirability, there was no indication of any underlying reason for the different response options selected. As such, this finding was probed further by asking "what would make you choose 0/+1/+2?" where the number inserted was an option that the participant did not select. This demonstrated that participants themselves were unclear as to which number was the most accurate reflection of their internally generated response for numerous items:

> *"I don't always know for the 0 and +1 and +2 where I should sit if I'm planning to keep the same treatment plans and I think I was probably inconsistent about that throughout the test."* – SRRES01

> *"I guess in my head I thought it's not a 0 because what I know now is different than what I knew when I went in…so it doesn't make my plan better…I don't have a good answer for why +1 and not 0. Does this happen to other people? I'm kind of like sure it's fine either way."* – EPSS01

> *"In this case, honestly the numbers are kind of confusing because I was planning to do it, and I'd still do it, so I don't know what number that is. I can't answer that question."* – SRRES05

> *"I would still do [the planned management], so +2. Um, actually no wait. See why are these numbers so hard?! Actually it changes nothing, it confirms what I knew. So 0. But it reaffirms it to the positive. So +? I'm still going do [the planned management]. Okay wait let me think of these numbers…I don't know! +1. Because it's in between. Good luck with your project. Am I insane?"* – EPSS01

Participants' difficulty determining what number corresponded with their desire to proceed with the planned management resulted in a range of response options. This was not because of intended normal variation in decision-making processes, but rather because of variable interpretation of the numerical scale. The intention is that if the new finding makes the planned management more indicated, the response selected should be +1 or +2 depending on the strength of the indication. However, if the new finding does not impact the planned management, the response selected should be 0 because of the planned management is neither more or less indicated than it initially was.

The next issue demonstrated with response matching originates from the rationale used to determine the strength of the indication or contraindication. In other words, why participants chose +/-1 versus +/-2. This issue is unique from the previously described issue which is only present when participants agree with the plan, and involves the 0 response option. Instead, here participants describe a variety of reasons that they use to determine whether 1 or 2 matches their internally generated response. As described previously, one such factor is social desirability. Other factors were observed:

> *"I guess it's the degree of what I had to do that corresponded to my number. If I'm just oversewing, it's a small intervention so -1. Bigger intervention [like a resection] is -2…what does -1 and -2 really mean? I don't know."* – EPSS01
>
> *"I would pick -2 because there are 2 factors making me think not to do this."* – SRRES01
>
> *"If I thought something was wrong then -2, but if I think something [in addition to the planned management] then -1."* – SRRES01
>
> *"It's difficult when there's 2 degrees of positive and 2 of negative. Would I kind of not do it or really not do it? I don't know. It's just no. Some questions makes sense to have in between, sometimes no is just no."* – EPSS01

Among the factors observed to influence participants' decisions were the magnitude of the intervention thought to be indicated, the number of reasons internally generated, whether something was inherently wrong versus insufficient, or no rationale at all. Interestingly, the rationale was not only noted to vary across participants, but within each participant (as above), the rationale for deciding whether something was just indicated/contraindicated versus strongly indicated/contraindicated changed across items. Various solutions have been proposed in existing literature discussing scoring issues with the SCT, but the optimal solution remains unknown. The various propositions are discussed further in the proceeding *Chapter* 4: Discussion chapter.

The notion of how to decide which response to pick when the planned management was thought to be insufficient, but not incorrect, led to the observation of the next response matching issue. In other words, participants sometimes agreed with the planned management but felt it was correct when performed in addition to another procedure. Participants variably interpreted whether this meant the planned management was correct because it was still part of their total management plan or incorrect because the planned management proposed it alone.

> *"Probably still do the [planned management], but investigate small bowel as well. It's tricky because ONLY [the planned management] is probably not appropriate, but the [planned management] doesn't change. If the question is whether you would do the [planned management] or not, then I'd pick 0 or +1 but if it's asking would I do only [the planned management] then my number would change toward the negative side."* – JRRES03

> *"I pick -1 here, because we still need [the planned management] but there should be consideration of [doing more]. I don't have a good way to express that, I'm limited to a few numbers. I agree with much of the plan proposed, but not all of it because I think something additional is needed…it's difficult to fit a complex answer into an ordinal scale."* – SRRES01

> *"0 is because you need additional management but [the planned management] isn't wrong."* – JRRES02

> *"Would you do [the planned management] AND something else as well? I think it's just a little ambiguous based on your interpretation of the question… just another layer is part of the question and I don't know how you add that in based on this format…I think I'd still say a +2, it's still indicated, but there's more you need to do."* – SRRES06

Although the issue of whether the question is asking if the planned management is still indicated or if only the planned management is indicated can be seen as a comprehension issue, it was coded as a response process issue because participants are clear on their internally generated answer, but are unclear on how to express it in the numerical format of the test question. Thus, the issue is not with respect to their comprehension of the question, but with the meaning behind the numerical responses when the planned management is part, but not all, of their internally generated management plan. A proposed solution to this problem is to incorporate the test-makers' intent in the instructions for the exam. Whether the planned management being part of the entire plan makes it indicated or not is relatively less important than ensuring consistency across all test-takers, including both trainees and the expert scoring panel.

The final issue noted with respect to response matching relates to the involvement of time in one item (Figure 7):

Scenario 15)

You have admitted a 67 year old male with a small bowel obstruction. After 24 hours the patient's symptoms have not been relieved. He continues to be obstipated, and have a tender and distended abdomen. Repeat abdominal films continue to show multiple loops of distended small bowel with several air fluid levels. You decide to take the patient to the operating room for an exploratory laparotomy. Given that the patient has had previous abdominal surgery for a perforated ulcer you expect that you will need to perform a simple lysis of adhesions.

If you were planning…	and in the OR found…	the planned management is:				
48) An exploratory laparotomy and lysis of adhesions	Just prior to induction you are told that his potassium 2.1	-2	-1	0	+1	+2

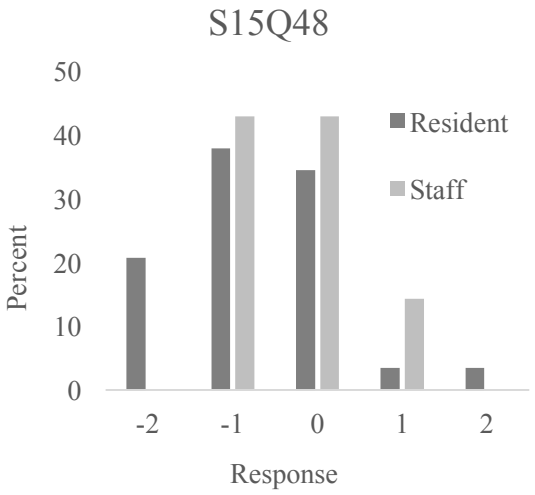

Figure 7. *Scenario 15, Question 48 with graph demonstrating resident and staff responses*

For this item, participants described wanting to pursue the planned management, but after a time delay to correct for the new finding. The time delay desired ranged from hours to days depending on the participant. This posed a response matching issue because participants variably interpreted indicated/strongly indicated to mean at this moment or at some point in the near future:

> *"I would probably…try and correct it then operate. So I would do the same management, but not at that time. So I would pick 0 because I'm still going to do it, but not now."* – JRRES03

> *"This patient isn't urgently unwell so you could probably correct K first. So I say -1 because they need to be optimized further… either way we need to do the planned management, it's just a matter of when…you could sit on it longer, but ya"* – JRRES02

> *"Does delaying the OR count as a -2? Depends how I interpret it? To me [time is] not [a factor] here. I would still proceed with the procedure, but delay it to correct the electrolyte imbalance."* – SRRES03

An added layer for this item were that participants did not necessarily agree with the planned management, as described in the comprehension category. As demonstrated in Figure 7, most participants reported the planned management to be contraindicated or strongly

contraindicated. Probing revealed that the most common rationale was because this was a non-urgent indication and the patient could easily wait. While modifying the question so that the patient urgently needed surgery would resolve the variable interpretation of the impact of time delay on response matching, it would also change the intent of the item, which may be to test the test-takers' ability to decide not to operate, which may require more experience to decide (Szatmary, Arora, & N, 2010). Unfortunately, any question in which the planned management is non-urgent poses the time delay issue, and as such another option would be to again specify the test-makers' intent in the SCT instructions. This proposed solution, however, may also be less than ideal as the cut-off of how much time corresponds to what response would likely also be variable.

In summary, the response matching cognitive process revealed the greatest number of issues with respect to the SCT's response process validity evidence. Since the numerical Likert-type scale is unique to the SCT, it is expected that the numerical scale generates unique issues. For some of these issues, simple solutions were proposed, such as clarifying the intended use of the scale in the SCT's instructions. For other issues, the solutions are complex as they would require major revisions to the numerical scale. Many of these issues have been described elsewhere in the literature as they pertain to SCT scoring issues. A summary of these issues and proposed solutions for future study are presented in Table 9, and are further reviewed in the *Chapter* 4: Discussion chapter of this thesis.

Table 9 *Findings of SCT Identified in Cognitive Interviews*

Category	**Description of Finding**	**Proposed Solution**
Comprehension	- Varying interpretation of term 'no intervention'	- Change planned management to an operative procedure
	- Desire for additional detail within item stem	- No change needed, clarify inherent ambiguity in SCT instructions

	- Disagreement with initial planned management	- Seek expert consensus on planned management of items
Recall	- Experts lack knowledge to answer subspecialized items outside their area of practice	- Remove subspecialized items for elective procedures or use subspecialty-specific experts for scoring panel
	- Trainees may not yet have learned the necessary knowledge to trigger recall process	- N/A
Decision-making Motivation	- Appropriate use of decision-making process to answer items	- N/A
Social desirability	- Desire to demonstrate humility vs. desire to appear confident	- Unknown
Response matching	- The meaning of 0	- Clarify intent of 0 in SCT instructions or omit 0 response option from all items
	- Which response corresponds with agreeing to planned management	- Specify in SCT instructions: if new finding has no impact = 0, if new finding makes more indicated = +1 or +2
	- Rationale for strength of (contra)indication	- Unknown
	- Desire to do planned management plus an additional procedure	- Specify in SCT instructions
	- Influence of time	- Remove non-urgent items or specify in SCT instructions

Chapter 4: Discussion

Overview

The findings of this study provide test content and response process validity evidence for the SCT. While the test content validity evidence gathered is specific to the SCT used in this study, the response process validity evidence may inform the design of other SCTs as well. This chapter explores the relationship between the findings of this study and the decision-making and validity theory presented in the *Introduction* chapter. The implications of the findings as they relate to existing literature and the use of the SCT will be examined. Finally, the limitations of this study and potential future research will also be discussed.

Conceptual Framework

The *Chapter* 1: Introduction chapter presented two models: the dual process theory of decision-making and Tourangeau's cognitive model of the response process when answering a question. The SCT is designed within the dual process framework of decision-making, and its intent is to measure the extent to which trainees' scripts (a System 1 non-analytic resource) are concordant with experts' scripts. To better understand test-takers' response processes, Tourangeau's cognitive model was used to analyze the results of the cognitive interviews in collecting response process validity evidence for the SCT.

The dual processing theory of decision-making describes the interplay of quick and accurate processing of recognized patterns through System 1 processing, and conscious and deliberate processing of unrecognized patterns for System 2 processing (Norman et al., 2017). While each process can function independently, the two processes can also oscillate with both systems influencing the ultimate decision (Goos et al., 2016). While it is difficult to definitively determine that System 1 versus System 2 processing has occurred, two possible indicators include response time and evidence of non-analytic resource use. As mentioned in the *Chapter* 1: Introduction chapter, Gagnon et al. (2006) demonstrated that response time for SCT items was significantly faster when presented with information typical of the hypothesis, compared to atypical information. This indirectly suggests that typical information activates recognized patterns which enables a greater degree of the faster System 1 processing (Gagnon et al., 2006). The response process validity evidence gathered in this study provides evidence for the use of non-analytic resources, such as scripts. As clinicians gain experience, they associate new problems with their past experiences, which results in the faster and more accurate decision-

making characteristic of System 1 processing (Brush et al., 2017; Moulton et al., 2007). In this study, senior trainees and staff frequently pointed to their relevant experience when making decisions, and often cited their experience as a prime determinant of their ultimate decision for cases of clinical uncertainty. Similarly, junior residents pointed to a lack of comfort making a decision on an issue they had not seen previously even when they felt they understood the relevant knowledge. This suggests the predominance of System 1 processing when the item activates the relevant scripts formed from previous experience, and the increased reliance on System 2 processing when the test-taker does not have clearly formed scripts from which to draw on.

Although the purpose of the SCT is to assess the concordance of expert and trainee scripts, the results of this study demonstrated that it is difficult to isolate and test solely these scripts. In the context of the SCT questions, cognitive interviews demonstrated that additional cognitive processing occurs. Tourangeau's cognitive model of the response process when answering a question clearly divides the process into four steps: comprehension, recall, decision-making, and response matching (Willis, 2015). The results of this study fit within Tourangeau's cognitive model by identifying discrete findings and/or issues within each stage of the model. The model proposes a straightforward sequence where each stage must be "passed" before the test-taker can move toward the next stage (Willis, 2015). For example, the decision-making stage cannot be triggered if the test-taker does not comprehend the item or is unable to recall the relevant knowledge from memory.

While the purpose of the SCT is to focus on the predominantly System 1 processing that occurs when scripts are activated by relevant experience, the results of this study demonstrate that noise within the comprehension, recall, and response matching cognitive steps exist within

the response process of test-takers. Evidence supporting the intended SCT purpose within the dual process theory and the noise observed within Tourangeau's cognitive model informed the conceptual framework for this thesis. The results of this study demonstrated that when answering a SCT question, the test-taker must first comprehend the item then recall the relevant information to trigger a decision-making process. Within the decision-making process, a combination of System 1 and System 2 processing is used depending on the extent to which the test-taker's experience results in pattern recognition, and thus the activation of relevant scripts. Once a decision has been made, the test-taker must then map their internally generated response to the response options provided by the item. A graphical depiction of this conceptual framework is presented in Figure 8.

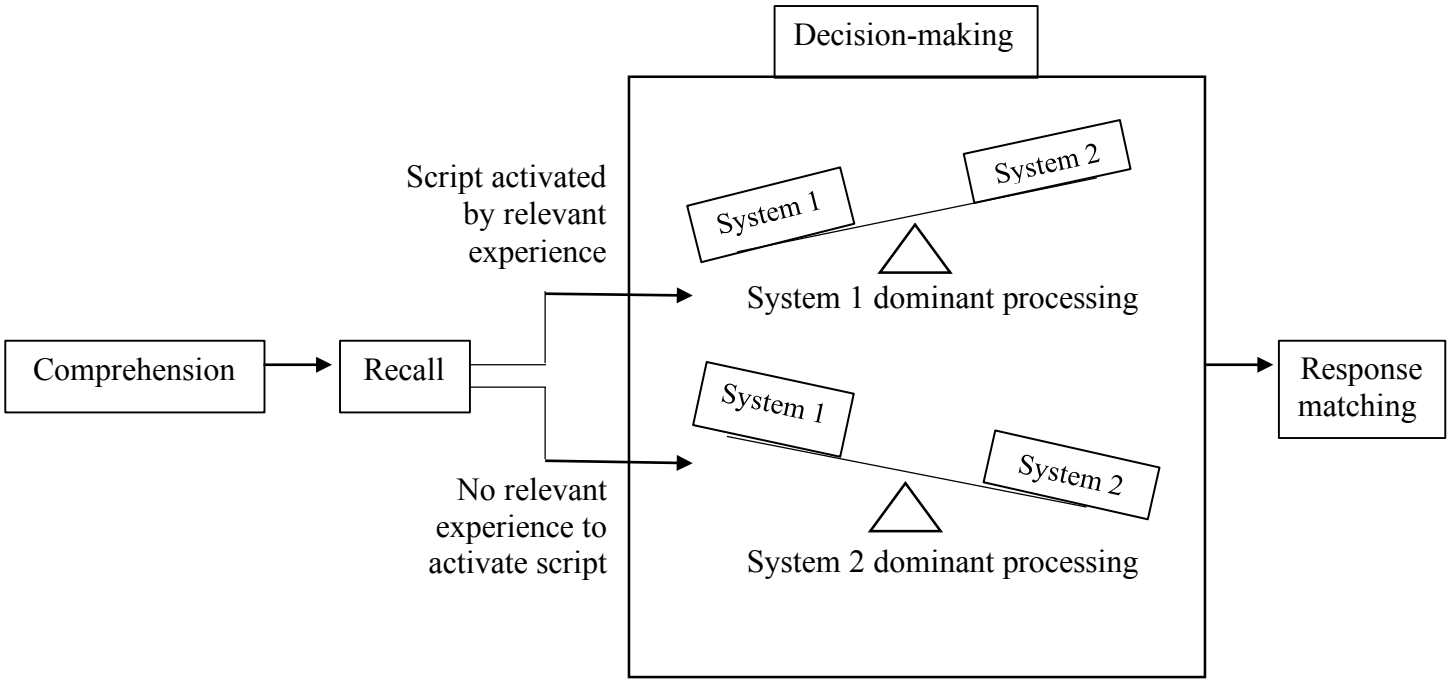

Figure 8. *Conceptual Framework Depicting the Cognitive Process when Answering a Script Concordance Test Question*

Content Validity Evidence

The purpose of the SCT is to assess CDM in cases of clinical uncertainty (Fournier et al., 2008). The results of the Delphi demonstrated that the SCT used in this study tests intraoperative decision-making in realistic and challenging general surgery cases of uncertainty, which is consistent with the intended purpose. However, within the scope of general surgery, both the original and final SCT more accurately represent a test of open, adult general surgery with several notable exceptions on the final SCT.

First, the absence of hepatobiliary cases from the final SCT was because the expert panel perceived the hepatobiliary cases to be too difficult, and more appropriate to the fellow level of training. This may be because the original SCT only included open cases, which tend to be performed open because of their complexity, while the simpler and more common procedures are often done via minimally invasive techniques. For example, the most common procedure done by general surgeons is a cholecystectomy (Fingar, Stocks, Weiss, & Steiner, 2012), which falls under the hepatobiliary competency, but was not included in any of the original SCT's cases. While the reason cholecystectomy cases were not included is unknown, it may be because the modern-day cholecystectomy is typically performed through minimally invasive techniques (Haribhakti & Mistry, 2015). Since the open hepatobiliary cases were thought by experts to be too complex for the resident level of training, and minimally invasive surgery is becoming the gold standard for a growing number of procedures (Buia, Stockhausen, & Hanisch, 2015), the inclusion of minimally invasive procedures should be considered in future general surgery SCTs to ensure representation of hepatobiliary surgery, and modern-day general surgery techniques as a whole, in the SCT.

The second notable absence from the final SCT was that of breast surgery cases. For these cases, their exclusion resulted from a perceived lack of realism, where experts felt that modern day pre-operative imaging rendered unexpected intraoperative findings extremely unlikely. This has important implications for all SCTs in that there may be certain subject areas with inherently minimal uncertainty, which therefore do not lend themselves well to the intended purpose of the SCT. This, however, does not necessarily mean that breast surgery cannot be tested via the SCT. Although breast surgery may have little intraoperative uncertainty, instead the uncertainty may lie in the diagnosis and operative planning stages, which may be more suitable for testing via a pre-operative SCT. Ensuring the subject area does carry inherent uncertainty is not included in the SCT construction guidelines (Fournier et al., 2008), but may be a relevant consideration when developing an SCT specific to one phase of care (i.e. intraoperative) for a given subspecialty (i.e. breast).

The final notable omission from the original to final SCT came from the exclusion of esophagus and head and neck competencies, which experts felt to be outside of the scope of practice for a modern day general surgeon despite the Royal College citing them as medical expert competencies of general surgery training (Royal College of Physicians and Surgeons of Canada, 2017). In academic practice, esophagus cases are typically seen by thoracic surgeons, while head and neck cases are seen by ear nose and throat (ENT) surgeons. However, in community practice, general surgeons do see minor esophagus and head & neck pathology, which may explain the discrepancy. Since our expert panel consists entirely of academic surgeons, this is an important limitation that may explain the discordance. However, since most residency training occurs within academic centers, the opinion of an academic expert panel may also more accurately reflect the exposure of general surgery trainees. The discordance between

expert and Royal College definitions of the scope of general surgery is imperative to clarify, particularly with the purpose of designing high-stakes examinations.

Response Process Validity Evidence

Prior to this study, response process validity of the SCT was weakly supported (Lubarsky et al., 2011). The cognitive interviews in this study provide some response process validity evidence for the SCT by demonstrating that to a large degree, experts and examinees interpret the SCT items in a manner consistent with the purpose of testing decision-making in cases of clinical uncertainty. However, investigating how experts and examinees choose their answers elucidated several issues with the SCT that interfered with test-takers' ability to answer questions in a manner that matches the intended purpose of the test. These issues were categorized into findings related to comprehension, recall, decision-making, and response matching (Willis, 2015). The previous chapter described these issues as well as proposed solutions. These issues include variable interpretation of terms and numerical response options, the desire for additional detail, and an inherent lack of knowledge. Many of the proposed solutions include modifications to the SCT instructions for added clarity. Perhaps of greater relevance than issues specific to the SCT used in this study, however, are the issues that may be shared by other SCTs.

SCT instructions.

Several of the solutions proposed focused on the instructions provided to test-takers. The instructions provided in the SCT explained the purpose and format of the SCT (Appendix 3) and provided an example case with accompanying questions, and a clear indication of the anchors for the Likert scale (Figure 1. *Example of an SCT Case (Fournier et al, 2008)*) as demonstrated in

the SCT Guidelines for Construction (Fournier et al., 2008). Of note, the anchors used for the Likert scale were the same for all questions on the SCT used in this study (strongly contraindicated, contraindicated, neither more or less indicated, indicated, strongly indicated). Despite the instructions provided, variable interpretation was noted with respect to several aspects of response matching including the meaning of 0, how to express agreement with the planned management on the numerical scale, how to express the desire to do the planned management plus an additional procedure, and the influence of time. Although the described issues have not previously been elucidated directly from test-takers' response process, this study is not the first time response matching issues have been proposed. For example, Kreiter (2012) used a Bayesian equation to demonstrate that when the planned management is certain, and the intraoperative finding adds no additional information, the test-taker must then decide between using the scale to indicate the diagnosis is certain or to indicate that the new finding adds no useful information. This theoretical explanation is supported by the finding that participants in this study were unclear on how to indicate their agreement with the planned management using the Likert scale. Literature exists on the SCT's theory (Charlin & van der Vleuten, 2004), construction (Fournier et al., 2008), number and content of cases and questions (Charlin et al., 2010), the expert panel (Gagnon et al., 2005), and scoring (Charlin et al., 2010), but few studies mention the instructions provided to test-takers (Sjoukje van den Broek et al., 2012). The described issues with respect to response process suggest that ensuring the test-makers' intentions are clearly described in the instructions is paramount to ensuring test-takers respond in a manner consistent with the intent.

Specialty-specific scoring.

With respect to the issue of recalling knowledge from long-term memory, cognitive interviews with expert participants suggested that they lacked the knowledge to answer subspecialized items outside their area of practice. This finding supports the conclusions described by Petrucci et al. (2013) that suggest using specialty-specific experts to develop the scoring key. Interestingly, the study by Petrucci et al. used the same original SCT that was used for the Delphi in this study. The rationale for using specialty-specific experts is that general surgery is a broad specialty encompassing many subspecialties. Accordingly, a surgical oncologist may not be a true expert in current trauma surgery practices (Petrucci et al., 2013). This poses a feasibility barrier, however, as SCT panels are recommended to consist of at least 10 experts to achieve acceptable reliability (Gagnon et al., 2005). But the number of surgeons representing each subspecialty within an institution is much smaller – for example, Ottawa currently has four hepatobiliary surgeons, four trauma surgeons, and four breast surgeons. In the context of summative assessment, it may therefore be necessary to employ surgeons from multiple institutions for each subspecialty. This adds another layer of complexity, however, as considerable variation in clinical practice exists between institutions (Bauer, 2009; Greenberg et al., 2011) which could result in scoring advantages or disadvantages depending on the examinee's institution of training relative to the institution of the expert scoring panellists. Consideration should be given to using specialty-specific experts when creating the scoring key for SCTs of broad specialties such as general surgery, but the potential consequences on test scores must be further studied before conclusions regarding this recommendation can be made.

Construct-irrelevant use of the numerical scale.

As with the above mentioned literature suggesting the use of specialty-specific experts for scoring (Petrucci et al., 2013), literature criticizing the SCT has hypothesized several other issues with respect to construct-irrelevant test taking strategies and alternative scoring methods. However, few studies demonstrate empirical evidence with respect to response processes to support or refute the proposed criticisms. The results of this study demonstrate evidence to support some of these hypotheses.

Since the SCT is based on ambiguous clinical scenarios, it is less likely for cases to be written that warrant the associated certainty of strongly refuting (-2) or strongly confirming (+2) a hypothesis (Lineberry et al., 2013). When written, cases with modal responses of +/-2 will be more factual and thus easier and less discriminating, resulting in these questions being more likely to be removed from an exam compared to questions with non-extreme modal responses (Lineberry et al., 2013). This further reduces the number of questions with extreme modal responses. In addition, when an extreme scale point is the modal response, non-modal responses can only pull credit toward the midpoint, as opposed to credit being pulled in either direction for non-extreme modal responses (-1, 0, or +1) (Lineberry et al., 2013). Because of all three phenomena, an examinee will score higher by consistently guessing the midpoint (-1 or +1) than by consistently guessing extremes. The results of this study demonstrate that test-takers themselves are not always clear on why they may pick moderate or extreme responses for a given question. Participants described being inconsistent, picking randomly, and citing construct-irrelevant methods of choosing a response such as basing their choice on the number of reasons for the decision or the magnitude of the intervention. The lack of a clearly defined rationale for choosing a moderate or extreme response combined with the construct-irrelevant advantage of

choosing moderately supports the literature demonstrating that certain racial or ethnic groups such as Hispanics and African Americans known to favour extreme responses may be disadvantaged compared to other ethnic examinees, such as Asians, whom are known to choose more moderately (Bachman & O'Malley, 1984; Marin, Gamba, & Marin, 1992; McDaniel, Psotka, Legree, Yost, & Weekley, 2011). Lineberry et al. (2013) take the implications of construct-irrelevant test-taking strategies one step further by suggesting that test-takers may quickly learn to significantly increase their scores by never choosing extreme options, and moreover, can often outperform the mean by simply selecting '0' for every question. The results of this study demonstrate that social desirability does impact the decision-making of trainees on the SCT, which makes the potential to select advantageous (i.e. moderate) response choices particularly concerning. Evidence demonstrating that test-takers' decisions are influenced by the desire to respond in a way that makes them look "better" suggests that with increasing familiarity to the SCT, test-takers may alter their responses to appear more moderate in their decisions as it could make their score look "better".

The issues elucidated with respect to response matching also relate to the extensive literature on the issues with the SCT's numerical scale. For example, Lineberry et al. (2013) argue that an examinee with perfect knowledge of experts' contradictory opinions about an item could reasonably surmise that choosing the '0' response option is the only way to convey his or her acknowledgement of the divided expert opinion. Participants in this study state exactly that: one reason for selecting 0 is because of the recognized clinical equipoise. However, such a participant could receive no credit for their response if the expert panel is in opposition on the item (i.e. half of experts choose -1 and the other choose +1). Perhaps more alarming is that along with choosing 0 to acknowledge clinical equipoise, participants in this study chose 0 to mean a

variety of things depending on the question. In general, if participants could not match their internally generated response with a numerical response option, ‘0’ appeared to be the default choice. Test-makers clearly describe that “The 0 anchor is not a shelter for candidates without a clear opinion, unlike the anchor "I don't know" in the Likert scale of an opinion poll” (Fournier et al., 2008). The results of this study demonstrate the test-takers do not necessarily respond to the 0 anchor in the way it is intended. The omission of ‘0’ was proposed by several study participants given the multiple ascribed meanings, but has not been studied in existing SCT literature. Omission of a neutral midpoint to compel test-takers to more clearly indicate their agreement or disagreement has been used in consensus methodology, such as the Delphi technique (de Villiers et al., 2005). Future studies may consider investigating the consequences of omitting the ‘0’ response option in terms of response process and scoring.

The SCT’s novel approach to CDM assessment yields several novel issues with respect to use of the Likert scale. Considering the thought process of test-takers when answering SCT questions is imperative to understanding the strengths and weaknesses of the unique test format. The significant issues elucidated with the numerical scale in this study suggest that further modifications are necessary prior to consideration in summative assessment.

The SCT for Formative Assessment

The purpose of this study was to investigate the use of the SCT for summative assessment. Although results demonstrated considerable test content and response process validity evidence supporting the SCT’s purpose of assessing decision-making skills in cases of uncertainty, the issues identified with respect to response process preclude the use of the SCT for

summative purposes. In its current form, the SCT may be better suited to formative assessment than summative.

As opposed to summative assessment, which is concerned with summarizing the achievement status of a student, formative assessment is concerned with how judgments about the quality of student responses can be used to shape and improve his or her competence. Feedback is a cornerstone of formative assessment, and can be defined as the information about the gap between a student's performance level and a reference level, which is used to decrease the gap in some way (Sadler, 1989).

Use of the SCT with the purpose of formative assessment has been studied (Cobb, Brown, Hammond, & Mossop, 2015; Gibot & Bollaert, 2008; Hornos et al., 2013). Participants in a study by Cobb et al. (2015) reported that the SCT encouraged them to reflect upon their clinical experience and participate in discussions of case material. Gibot et al. (2008) suggested administering the SCT early in training to identify trainees with difficulties in CDM. Hornos et al. (2013) allowed test-takers to view the expert panel's responses and justifications for their answers to stimulate reflection among learners. Although not formally investigated within this study, anecdotally the trainees and experts in this study described the SCT as a useful and novel method of assessment and several have requested to use the SCT to stimulate discussion around difficult cases in preparation for the Royal College oral exams.

Giving trainees the opportunity to provide written explanation to support their numerical responses, or to verbally discuss their responses with an examiner such as during the cognitive interview, may facilitate feedback between experts and trainees. Furthermore, written explanation or verbal discussion may alleviate the impact of issues related to the numerical scale

and scoring method to some degree, and instead shift the focus of the cognitive response process back to its intended purpose: the assessment of dual process decision-making.

As such, although the results of this study preclude supporting the SCT for summative assessment, the existing literature on use of the SCT for formative assessment and the validity evidence gathered in this study suggest that the SCT may be suitable for the formative assessment of CDM among surgical trainees.

Limitations

The results of this study should be interpreted with the understanding of several limitations. First, there is no evidence-based definition of experts for the Delphi technique and the SCT panel (Norman, 2005). Consensus literature describes a suitable expert as "someone who possesses the relevant knowledge and experience and whose opinions are respected by fellow workers in their field" (de Villiers et al., 2005). Similarly, SCT literature recommends that expert "selection decisions reflect accepted community standards of expertise in a given field" (Lubarsky, Dory, Duggan, Gagnon, & Charlin, 2013). Suggestions include formal certification in a field, a pre-specified number of years in practice, or an established reputation for sound clinical acumen (Lubarsky, Dory, et al., 2013). This study sought to use staff surgeons thought to be the top decision-makers by their colleagues to form the Delphi and SCT expert panels given the focus of assessing decision-making skills. To identify these surgeons, a survey was sent to all resident and staff surgeons within our division. The opinions of staff surgeons were solicited as they were thought to have a better understanding of the established reputation of their colleagues, while the opinions of resident surgeons were solicited as they were thought to value the opinions of the staff surgeons whose decision-making they hold in the highest regard. However, the

composition of the expert panel may have been different, and therefore potentially led to different results, had the expert panel instead been constructed based on other parameters such as years in practice, or the opinion of only staff surgeons.

The expert panel is limited also in that it was comprised only of surgeons practicing at The Ottawa Hospital. The local nature serves as a limitation to the generalizability of the entire study as well. It is unlikely that elucidated response process issues, such as the variable use of the '0' anchor, are unique to our institution, but the answers given by staff and resident surgeons for the Delphi and cognitive interviews are likely shaped by the culture of practice within our institution. For example, as described under

Results: Test Content Validity, the results of the Delphi technique may have included Head & Neck SCT items if the expert panel had consisted of community surgeons who routinely see Head & Neck cases as part of their clinical practice. Similarly, the response patterns of experts and residents on the final SCT may have favoured operative intervention at an institution that routinely operates on diverticulitis, as opposed to the predominantly non-operative approach practiced in Ottawa. The study was conducted locally for several reasons. In addition to the improved feasibility of a locally conducted study, it is important that the expert and resident test-takers represent the same practice groups. As in the examples above, recruiting community surgeons outside of the hospitals in which the resident test-takers train, or recruiting expert and resident participants from several different institutions may introduce noise where the response process of each group is a result of variations in cultural practice as opposed to variation in individuals' CDM. However, variations in practice are an important part of the SCT, which is founded on the principle of multiple correct answers (Goos et al., 2016). As such, it is less important where the experts practice and more so that the institution in which the experts practice matches the institution in which the residents train. If designing an SCT with the

purpose of high-stakes examination, it would be important to recruit a national expert panel with similar composition to the resident examinees.

The remaining limitations are unique to the cognitive interviewing portion of this study. First, all the interviews were conducted by the principal investigator, who is a general surgery resident being trained at the institution of study. Thus, a pre-existing relationship existed between all the participants and the interviewer which may have influenced both the test-takers' responses during the cognitive interviews as well as the data analysis. To understand the content of responses, it is necessary to have a subject matter expert conduct the interviews, but this inevitably results in bias given the social hierarchy that exists within the culture of surgery. Given most participants were in a position of power over the researcher, this may have enabled participants to be at ease more so than if the researcher was in a position of power relative to the participants. As described in the *Methods* section, this limitation was mitigated by various steps including prospective selection of unbiased and deductive probes, reflexivity of the research team, and a second independent analysis of the data by a non-surgeon research assistant. Transcripts were also reviewed by a second subject matter expert to ensure there was no evidence of discomfort or dishonesty from participants.

The next limitation with respect to the cognitive interviews also stems from the interviewee's responses. During data analysis, it was noticed that despite one or more interviewees selecting a given response option during the cognitive interview, no participants had selected that response option during administration of the final SCT. This observation was corroborated by statements made by interviewees acknowledging that perhaps their responses selected during the interview were not the same as the responses selected during test administration. Although not quantified, possible reasons for the suboptimal intra-rater reliability

include that the cognitive process during the interview is different than during test administration, social desirability during the cognitive interview altered interviewee responses, or response matching issues (i.e. the meaning of 0, rationale for strength of (contra)indication) led to some degree of arbitrary response matching. While it is likely that all reasons listed contribute to some extent, after resolving the described response process issues, future studies could consider concurrent probing where the interview probes are applied as the questions are tested (Willis, 2015) to identify and mitigate the question of intra-rater reliability. The issue of intra-rater reliability is particularly important when considering intra-panellist reliability, as the SCT scoring key is dependent upon reliable responses from each panellist and this has not been purposefully studied (Lineberry et al., 2013).

The final limitation with respect to the cognitive interviews stems from the sampling of SCT cases and items to be probed during the interview. Due to the significant cognitive load and time required to answer the SCT, the additional cognitive load and time required to delve into the cognitive process for each item precluded the inclusion of all cases and items. The graphically depicted results of the administered final SCT were used by two subject matter experts to decide which cases and items to probe in the cognitive interviews. This was done by selecting a range of response patterns, for example some cases/items with a high degree of expert panel agreement and resident participant agreement, some with a high degree of expert panel agreement but poor resident agreement, some with poor expert panel agreement and poor resident agreement, etc. Nonetheless, it is possible that response process issues exist in cases/items that were not included in the cognitive interview and thus were missed by the methods used in this study. The results of this study did yield a considerable number of issues that are relevant to the construction of all

SCT cases, however, and thus future studies resolving the elucidated issues could aim to re-interview SCT-takers to determine whether unidentified issues exist or not.

Future Directions

Given little previous work in collecting response process evidence for the SCT (Lubarsky et al., 2011), the results of this study inform several avenues for future research by exploring modifications to the SCT for research and practical purposes. The first area pertains to the instructions provided to test-takers. As discussed previously, few studies address the instructions that have been or should be provided. The response process issues uncovered in the results of this study provide suggestions on sources of confusion and varying interpretation which may be resolved by explicitly clarifying test-makers' intent in the SCT instructions. Specifically, a consistent interpretation of each response process issue uncovered in this and other studies could be explicitly specified. For example, "with respect to time, if you want to delay the planned management to pursue interim management, consider the planned management contraindicated (-1) or strongly contraindicated (-2)". Future studies may investigate whether this resolves some of the response process issues described in this study or if more extensive modifications to the SCT are required.

Another area that warrants further investigation is the numerical scale and scoring methods of the SCT. Although the scoring methods were beyond the scope of this study, the use of the numerical scale directly relates to scoring methods through issues such as test-taking strategies. Due to the numerous issues with respect to the numerical scale, further refinement is necessary if the SCT is to be considered for summative assessment. Several studies have explored alternative response formats (Bland et al., 2005; Kelly, Durning, & Denton, 2012;

Petrella & Davis, 2007). Proposed modifications to the Likert scale have not considered the perspective of test-takers' response process, however. The results of this study demonstrate evidence of construct-irrelevant use of the Likert scale, particularly with respect to rationale for choosing the strength of (contra)indication (i.e. +2 vs. +1, -2 vs. -1). This may serve as further impetus to consider modifications to the Likert scale. For example, if test-makers want to maintain the greater level of discrimination offered by the 5-point scale, perhaps accompanying each question with free-text justification of answers (Kelly et al., 2012) could be considered to allow test-takers to explain their rationale for their response choice to ensure it is construct-relevant. Future study may also pursue purposeful investigation of the potential benefits and/or consequences of affording test-takers the opportunity to explain themselves in some format. Another possible option may to be to eliminate the Likert scale entirely and instead ask a series of forced-choice questions such as "would you do Procedure X given the new finding?", "do the patient factors in the case push you toward doing/not doing Procedure X?", and "does the new intraoperative finding push you toward doing/not doing Procedure X?". Although this modification has not been studied, in theory it would allow a clear understanding of which direction each factor pushes the test-taker as well as the combined weight of those factors to do or not to do Procedure X (i.e. the planned management).

Finally, the purpose of this study was to collect content and response process validity evidence for the SCT. However, the unitary view of validity describes validity as a single entity for which various sources are gathered into an integrated evaluative judgment, as opposed to separate kinds of validity (Messick, 1989). While test content and response process validity evidence are typically gathered during test development, the other three sources of validity evidence (internal structure, relations to other variables, and consequences of testing) are

typically collected during pilot testing or after test administration. Existing literature has made some attempts at the latter three sources of validity evidence as described in the *Chapter* 1: Introduction, but further investigation is warranted, particularly after modifying the SCT based on the response process validity issues uncovered by this study.

Conclusions

Assessment of competence in surgical training must include the purposeful assessment of decision-making skills. The Script Concordance Test offers a novel approach to CDM assessment that incorporates the realities of modern practice: many situations involve clinical uncertainty and there can be more than one correct answer. The use of the SCT for assessment should be based on a solid foundation of valid results supporting its intended purpose. Despite the extensive literature published on the SCT, notable gaps in validity evidence informed the research questions that guided this study. Using an SCT previously administered nationally, test content and response process validity evidence were gathered. While the test content validity evidence is more specific to the SCT used in this study, the response process validity evidence may be generalizable to other SCTs.

The test content validity evidence demonstrated that experts do generally agree that the SCT represents realistic and challenging general surgery cases that test clinical decision-making in situations of clinical uncertainty. Notable subspecialties that were excluded from the final SCT serve as a caution to test-makers to ensure items are an appropriate level of difficulty for the intended test-takers, and that some specific subjects (i.e. intraoperative breast surgery) may not lend themselves well to the SCT as they inherently are less likely to contain clinical uncertainty. The response process validity evidence used cognitive interviews to demonstrate that test-takers

interpret SCT items in a manner consistent with the intended purpose of assessing CDM. Investigating how staff and resident surgeons choose their answers revealed several issues with the SCT, particularly with comprehension and response matching stages of the cognitive process. While some of these issues likely have straightforward solutions such as clarifying the test-makers' intent in the SCT instructions, other issues are intrinsic to the SCT format and may require greater modifications to resolve them. As such, the results of this study support the use of the SCT for the assessment of CDM in surgical trainees, but cannot support its use for summative assessment in its current form given the elucidated issues. Consideration should be given to using the SCT for formative assessment. Further revision to the rating scale, followed by further collection of validity evidence, is necessary before the SCT can be used in summative assessment.

References

AERA, APA, & NCME. (2014). The Standards for Educational and Psychological Testing. Retrieved from http://www.apa.org/science/programs/testing/standards.aspx

Aggarwal, R., & Darzi, A. (2006). Technical-Skills Training in the 21st Century. *New England Journal of Medicine*, *355*(25), 2695–2696. http://doi.org/10.1056/NEJMe068179

Angelos, P. (2017). The right choice? Surgeons, confidence, and humility. *ACS Surgery News*. Parsippany. Retrieved from https://www.mdedge.com/acssurgerynews/article/130301/practice-management/right-choice-surgeons-confidence-and-humility

Bachman, J. G., & O'Malley, P. M. (1984). Yea-saying, nay-saying, and going to extremes: black-white differences in response styles. *Public Opinion Quarterly*, *48*(2), 491. http://doi.org/10.1086/268845

Bauer, V. P. (2009). Emergency management of diverticulitis. *Clinics in Colon and Rectal Surgery*, *22*(3), 161–168. http://doi.org/10.1055/s-0029-1236160

Bland, A. C., Kreiter, C. D., & Gordon, J. A. (2005). The psychometric properties of five scoring methods applied to the script concordance test. *Academic Medicine : Journal of the Association of American Medical Colleges*, *80*(4), 395–9. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/15793026

Brailovsky, C., Charlin, B., Beausoleil, S., Coté, S., & Van Der Vleuten, C. (2001). Measurement of clinical reflective capacity early in training as a predictor of clinical reasoning performance at the end of residency: An experimental study on the script concordance test. *Medical Education*, *35*(5), 430–436. http://doi.org/10.1046/j.1365-2923.2001.00911.x

Brailovsky, C., & Grand'Maison, P. (2000). Using Evidence to Improve Evaluation: A Comprehensive Psychometric Assessment of a SP-Based OSCE Licensing Examination. *Advances in Health Sciences Education : Theory and Practice*, *5*(3), 207–219. http://doi.org/10.1023/A:1009869328173

Brush, J. E., Sherbino, J., & Norman, G. R. (2017). How expert clinicians intuitively recognize a medical diagnosis. *American Journal of Medicine*, *130*(6), 629–634. http://doi.org/10.1016/j.amjmed.2017.01.045

Buia, A., Stockhausen, F., & Hanisch, E. (2015). Laparoscopic surgery: a qualified systematic review. *World Journal of Methodology*, *5*(54), 238–254. http://doi.org/10.5662/wjm.v5.i4.238

Carrière, B., Gagnon, R., Charlin, B., Downing, S., & Bordage, G. (2009). Assessing Clinical Reasoning in Pediatric Emergency Medicine: Validity Evidence for a Script Concordance Test. *Annals of Emergency Medicine*, *53*(5), 647–652. http://doi.org/10.1016/j.annemergmed.2008.07.024

Charlin, B., Boshuizen, H. P. A., Custers, E. J., & Feltovich, P. J. (2007). Scripts and clinical reasoning. *Medical Education*, *41*, 1178–1184. http://doi.org/10.1111/j.1365-2923.2007.02924.x

Charlin, B., Brailovsky, C., Leduc, C., & Blouin, D. (1998). The diagnostic script questionnaire: a new tool to assess a specific dimension of clinical competence. *Advances in Health Sciences Education*, *3*(1), 51–58. http://doi.org/10.1023/A:1009741430850

Charlin, B., Desaulniers, M., Gagnon, R., Blouin, D., & van der Vleuten, C. (2002). Comparison of an aggregate scoring method with a consensus scoring method in a measure of clinical

reasoning capacity. *Teaching and Learning in Medicine*, *14*(3), 150–156. http://doi.org/10.1207/S15328015TLM1403_3

Charlin, B., Gagnon, R., Lubarsky, S., Lambert, C., Meterissian, S., Chalk, C., … van der Vleuten, C. (2010). Assessment in the context of uncertainty using the script concordance test: more meaning for scores. *Teaching and Learning in Medicine*, *22*(3), 180–186. http://doi.org/10.1080/10401334.2010.488197

Charlin, B., Gagnon, R., Pelletier, J., Coletti, M., Abi-Rizk, G., Nasr, C., … van der Vleuten, C. (2006). Assessment of clinical reasoning in the context of uncertainty: the effect of variability within the reference panel. *Medical Education*, *40*(9), 848–854. http://doi.org/10.1111/j.1365-2929.2006.02541.x

Charlin, B., Gagnon, R., Sauvé, E., & Coletti, M. (2007). Composition of the panel of reference for concordance tests: do teaching functions have an impact on examinees' ranks and absolute scores? *Medical Teacher*, *29*(1), 49–53. http://doi.org/10.1080/01421590601032427

Charlin, B., Tardif, J., & Boshuizen, H. P. (2000). Scripts and medical diagnostic knowledge: theory and applications for clinical reasoning instruction and research. *Academic Medicine : Journal of the Association of American Medical Colleges*, *75*(2), 182–90. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10693854

Charlin, B., & van der Vleuten, C. (2004). Standardized assessment of reasoning in contexts of uncertainty: the script concordance approach. *Evaluation & the Health Professions*, *27*(3), 304–19. http://doi.org/10.1177/0163278704267043

Cobb, K. A., Brown, G., Hammond, R., & Mossop, L. H. (2015). Students' perceptions of the script concordance test and its impact on their learning behavior: a mixed methods study. *Journal of Veterinary Medical Education*, *42*(1), 45–52. http://doi.org/10.3138/jvme.0514-057R1

Collard, A., Gelaes, S., Vanbelle, S., Bredart, S., Defraigne, J.-O., Boniver, J., & Bourguignon, J.-P. (2009). Reasoning versus knowledge retention and ascertainment throughout a problem-based learning curriculum. *Medical Education*, *43*(9), 854–865. http://doi.org/10.1111/j.1365-2923.2009.03410.x

Cooke, S., & Lemay, J.-F. (2017). Transforming medical assessment: integrating uncertainty into the evaluation of clinical reasoning in medical education. *Academic Medicine*, *92*(6), 746–51. http://doi.org/10.1097/ACM.0000000000001559

Cronbach, L. (1971). Test Validation. In R. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443–507). Washington DC: American Council on Education.

Cuhls, K. (2005). Delphi method. Retrieved from http://www.unido.org.proxy.bib.uottawa.ca/fileadmin/import/16959_DelphiMethod.pdf

de Villiers, M. R., de Villiers, P. J. T., & Kent, A. P. (2005). The Delphi technique in health sciences education research. *Medical Teacher*, *27*(7), 639–643. http://doi.org/10.1080/13611260500069947

de Vries, E. N., Ramrattan, M. A., Smorenburg, S. M., Gouma, D. J., & Boermeester, M. A. (2008). The incidence and nature of in-hospital adverse events: a systematic review. *Quality & Safety in Health Care*, *17*(3), 216–23. http://doi.org/10.1136/qshc.2007.023622

Devlin, J. W., Marquis, F., Riker, R. R., Robbins, T., Garpestad, E., Fong, J. J., … Skrobik, Y. (2008). Combined didactic and scenario-based education improves the ability of intensive care unit staff to recognize delirium at the bedside. *Critical Care*, *12*(1), R19. http://doi.org/10.1186/cc6793

Dory, V., Gagnon, R., Vanpee, D., & Charlin, B. (2012). How to construct and implement script concordance tests: insights from a systematic review. *Medical Education*, *46*(6), 552–563. http://doi.org/10.1111/j.1365-2923.2011.04211.x

Epstein, R. M., & Hundert, E. M. (2002). Defining and Assessing Professional Competence. *Jama*, *287*(2), 226–235. http://doi.org/10.1001/jama.287.2.226

Fabri, P. J., & Zayas-Castro, J. L. (2008). Human error, not communication and systems, underlies surgical complications. *Surgery*, *144*(4), 557-63; discussion 563–5. http://doi.org/10.1016/j.surg.2008.06.011

Feltovich, P., & Barrows, H. (1984). Issues of generality in medical problem solving. In H. Schmidt & M. De Volder (Eds.), *Tutorials in Problem-Based Learning: A New Direction in Teaching the Health Professions*. The Netherlands: Assen.

Fingar, K. R., Stocks, C., Weiss, A. J., & Steiner, C. A. (2012). Most frequent operating room procedures performed in U.S. hospitals, 2003-2012. Retrieved June 19, 2018, from https://www.hcup-us.ahrq.gov/reports/statbriefs/sb186-Operating-Room-Procedures-United-States-2012.jsp

Finlay, L. (2002). Negotiating the swamp: the opportunity and challenge of reflexivity in research practice. *Qualitative Research*, *2*(2), 209–230. Retrieved from https://www.utsc.utoronto.ca/~kmacd/IDSC10/Readings/Positionality/reflex-2.pdf

Fives, H., & Didonato-Barnes, N. (2013). Classroom test construction: the power of a table of specifications. *Practical Assessment, Research & Evaluation*, *18*(3), 1–7. Retrieved from http://pareonline.net/pdf/v18n3.pdf

Flin, R., Youngson, G., & Yule, S. (2007). How do surgeons make intraoperative decisions? *Quality & Safety in Health Care*, *16*(3), 235–9. http://doi.org/10.1136/qshc.2006.020743

Fournier, J. P., Demeester, A., & Charlin, B. (2008). Script concordance tests: guidelines for construction. *BMC Medical Informatics and Decision Making*, *8*, 18. http://doi.org/10.1186/1472-6947-8-18

Fournier, J., Thiercelin, D., Pulcini, C., Alunni-Perret, V., Gilbert, E., Minguet, J., & Bertrand, F. (2006). Clinical reasoning assessment in emergency medicine: script concordance tests are more efficient to detect clinical experience than rich-context multiple-choice questions. *Pedagogie Medicale*, *7*, 20–30. Retrieved from https://scholar.google.ca/citations?view_op=view_citation&continue=/scholar%3Fq%3DClinical%2Breasoning%2Bassessment%2Bin%2Bemergency%2Bmedicine:%2Bscript%2Bconcordance%2Btests%2Bare%2Bmore%2Befficient%2Bto%2Bdetect%2Bclinical%2Bexperience%2Bthan%2Brich-co

Gagnon, R., Charlin, B., Coletti, M., Sauve, E., & van der Vleuten, C. (2005). Assessment in the context of uncertainty: how many members are needed on the panel of reference of a script concordance test? *Medical Education*, *39*(3), 284–291. http://doi.org/10.1111/j.1365-2929.2005.02092.x

Gagnon, R., Charlin, B., Roy, L., St-Martin, M., Sauvé, É., Boshuizen, H. P. A., & Van Der Vleuten, C. (2006). The cognitive validity of the script concordance test: a processing time study. *Teaching and Learning in Medicine*, *18*(1), 22–27. Retrieved from https://journals-scholarsportal-info.proxy.bib.uottawa.ca/pdf/10401334/v18i0001/22_tcvotsctapts.xml

Gawande, A. A., Zinner, M. J., Studdert, D. M., & Brennan, T. A. (2003). Analysis of errors reported by surgeons at three teaching hospitals. *Surgery*, *133*(6), 614–621. Retrieved from https://journals-scholarsportal-info.proxy.bib.uottawa.ca/pdf/00396060/v133i0006/614_aoerbsatth.xml

Gay, S., & McKinley, R. K. (2017). 'When I say… dual-processing theory': evidence, not assertion. *Medical Education*, *51*(10), 1086. http://doi.org/10.1111/medu.13351

Gibot, S., & Bollaert, P.-E. (2008). The script concordance test as a formative assessment tool in critical care medicine. *Pédagogie Médicale*, *9*(1), 7–18. http://doi.org/10.1051/pmed:2008037

Godellas, C. V., Hauge, L. S., & Huang, R. (2000). Factors Affecting Improvement on the American Board of Surgery In-Training Exam (ABSITE). *Journal of Surgical Research*, *91*(1), 1–4. http://doi.org/10.1006/jsre.2000.5852

Gofton, W. T., Dudek, N. L., Wood, T. J., Balaa, F., & Hamstra, S. J. (2012). The Ottawa Surgical Competency Operating Room Evaluation (O-SCORE). *Academic Medicine*, *87*(10), 1401–1407. http://doi.org/10.1097/ACM.0b013e3182677805

Goos, M., Schubach, F., Seifert, G., & Boeker Martin. (2016). Validation of undergraduate medical student script concordance test (SCT) scores on the clinical assessment of the acute abdomen. *BMC Surgery*, *16*(57). http://doi.org/10.1007/BF02289746

Graham, B., Regehr, G., & Wright, J. G. (2003). Delphi as a method to establish consensus for diagnostic criteria. *Journal of Clinical Epidemiology*, *56*(12), 1150–1156. http://doi.org/10.1016/S0895-4356(03)00211-7

Greenberg, C. C., Lipsitz, S. R., Hughes, M. E., Edge, S. B., Theriault, R., Wilson, J. L., … Lee, H. (2011). Institutional variation in the surgical treatment of breast cancer: a study of the NCCN. *Annals of Surgery*, *254*(2), 339–345. http://doi.org/10.1097/SLA.0b013e3182263bb0

Groves, M., Scott, I., & Alexander, H. (2002). Assessing clinical reasoning: a method to monitor its development in a PBL curriculum. *Medical Teacher*, *24*(5), 507–515. http://doi.org/10.1080/01421590220145743

Hall, J. C., Ellis, C., & Hamdorf, J. (2003). Surgeons and cognitive processes. *The British Journal of Surgery*, *90*(1), 10–6. http://doi.org/10.1002/bjs.4020

Handfield-Jones, R., Brown, J. B., Rainsberry, P., & Brailovsky, C. A. (1996). Certification Examination of the College of Family Physicians of Canada. Part 2. Conduct and general performance. *Canadian Family Physician Medecin de Famille Canadien*, *42*, 1188–95. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8704495

Haribhakti, S. P., & Mistry, J. H. (2015). Techniques of laparoscopic cholecystectomy: nomenclature and selection. *Journal of Minimal Access Surgery*, *11*(2), 113–8. http://doi.org/10.4103/0972-9941.140220

Hawkins, R. E., Margolis, M. J., Durning, S. J., & Norcini, J. J. (2010). Constructing a validity argument for the mini-Clinical Evaluation Exercise: a review of the research. *Academic Medicine : Journal of the Association of American Medical Colleges*, *85*(9), 1453–61. http://doi.org/10.1097/ACM.0b013e3181eac3e6

Holmboe, E. (2008). Practice Audit, Medical Record Review, and Chart-Stimulated Recall. In E. Holmboe & R. Hawkins (Eds.), *Practical guide to the evaluation of competence* (pp. 60–74). Philadelphia: Mosby Elsevier. Retrieved from http://docplayer.net/4261593-Practice-audit-medical-record-review-and-chart-stimulated-recall.html

Hornos, E. H., Pleguezuelos, E. M., Brailovsky, C. A., Harillo, L. D., Dory, V., & Charlin, B. (2013). The practicum script concordance test: an online continuing professional development format to foster reflection on clinical practice. *Journal of Continuing Education in the Health Professions*, *33*(1), 59–66. Retrieved from https://journals-scholarsportal-info.proxy.bib.uottawa.ca/pdf/08941912/v33i0001/59_tpsctatfrocp.xml

Hsu, C.-C. (2007). The Delphi Technique: Making Sense Of Consensus. *Practical Assessment, Research & Evaluation*, *12*(10), 1–8.

Humphrey-Murto, S., Varpio, L., Gonsalves, C., & Wood, T. J. (2017). Using consensus group methods such as Delphi and Nominal Group in medical education research Ã. *Medical Teacher*, *39*(1), 14–19. http://doi.org/10.1080/0142159X.2017.1245856

Iravani, K., Amini, M., Doostkam, A., & Dehbozorgian, M. (2016). The validity and reliability of script concordance test in otolaryngology residency training. *Journal of Advances in Medical Education & Professionalism*, *4*(2), 93–6. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/27104204

Kane, M. (2006). Validation. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport: Praeger.

Kelly, W., Durning, S., & Denton, G. (2012). Comparing a script concordance examination to a multiple-choice examination on a core internal medicine clerkship. *Teaching and Learning in Medicine*, *24*(3), 187–193. http://doi.org/10.1080/10401334.2012.692239

Kreiter, C. D., & Kreiter, C. D. (2012). Commentary: the response process validity of a script concordance test item. *Advances in Health Science Education*, *17*, 7–9. http://doi.org/10.1007/s10459-011-9325-0

Labelle, M., Beaulieu, M., Paquette, D., Fournier, C., Bessette, L., Choquette, D., … Thivierge, R. L. (2004). An integrated approach to improving appropriate use of anti-inflammatory medication in the treatment of osteoarthritis in Qu?bec (Canada): the CURATA model. *Medical Teacher*, *26*(5), 463–470. http://doi.org/10.1080/0142159042000218669

Lambert, C., Gagnon, R., Nguyen, D., & Charlin, B. (2009). The script concordance test in radiation oncology: validation study of a new tool to assess clinical reasoning. *Radiation Oncology (London, England)*, *4*, 7. http://doi.org/10.1186/1748-717X-4-7

Lane, S. (2014). Validity evidence based on testing consequences. *Psicothema*, *26*(1), 127–35. http://doi.org/10.7334/psicothema2013.258

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic Inquiry*. Newbury Park: Sage Publications. Retrieved from https://books.google.ca/books?hl=en&lr=&id=2oA9aWlNeooC&oi=fnd&pg=PA7&dq=lincoln+guba&ots=0tmtY7SfCl&sig=EDpm5ae2v237OBjuZnrbMv7qVW8#v=onepage&q=lincoln guba&f=false

Lineberry, M., Kreiter, C. D., & Bordage, G. (2013). Threats to validity in the use and interpretation of script concordance test scores. *Medical Education*, *47*(12), 1175–1183. http://doi.org/10.1111/medu.12283

Lubarsky, S., Charlin, B., Cook, D. A., Chalk, C., Cees, &, & Van Der Vleuten, P. M. (2011). Script concordance testing: a review of published validity evidence. *Medical Education*, *45*, 329–38. http://doi.org/10.1111/j.1365-2923.2010.03863.x

Lubarsky, S., Dory, V. R., Duggan, P., Gagnon, R., & Charlin, B. (2013). Script concordance testing: from theory to practice: AMEE Guide No. 75. *Medical Teacher*, *35*(3), 184–193. http://doi.org/10.3109/0142159X.2013.760036

Lubarsky, S., Gagnon, R., & Charlin, B. (2013). Scoring the script concordance test: Not a black and white issue. *Medical Education*, *47*(12), 1159–1161. http://doi.org/10.1111/medu.12362

Marin, G., Gamba, R. J., & Marin, B. V. (1992). Extreme response style and acquiescence among hispanics. *Journal of Cross-Cultural Psychology*, *23*(4), 498–509. http://doi.org/10.1177/0022022192234006

Marshall, M. N. (1996). Sampling for qualitative research. *Family Practice*, *13*(6), 522–5.

Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9023528
McDaniel, M. A., Psotka, J., Legree, P. J., Yost, A. P., & Weekley, J. A. (2011). Toward an understanding of situational judgment item validity and group differences. *Journal of Applied Psychology*, *96*(2), 327–336. http://doi.org/10.1037/a0021983
McDermid, F., Peters, K., Jackson, D., & Daly, J. (2014). Conducting qualitative research in the context of pre-existing peer and collegial relationships. *Nurse Researcher*, *21*(5), 28–33. http://doi.org/10.7748/nr.21.5.28.e1232
Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: Macmillan.
Meterissian, S. H. (2006). A novel method of assessing clinical reasoning in surgical residents. *Surgical Innovation*, *13*(2), 115–119. http://doi.org/10.1177/1553350606291042
Moorthy, K., Munz, Y., Sarker, S. K., & Darzi, A. (2003). Objective assessment of technical skills in surgery, *327*(7422), 1032–1037. http://doi.org/10.1136/bmj.327.7422.1032
Moulton, C.-A. E., Regehr, G., Mylopoulos, M., & MacRae, H. M. (2007). Slowing down when you should: a new model of expert judgment. *Acad Med*, *82*(10), 109–16. http://doi.org/10.1097/ACM.0b013e3181405a76
Newman, I., Lim, J., & Pineda, F. (2013). Content Validity Using a Mixed Methods Approach: Its Application and Development Through the Use of a Table of Specifications Methodology. *Journal of Mixed Methods Research*, *7*(3), 243–260. http://doi.org/10.1177/1558689813476922
Norman, G. (2005). Research in clinical reasoning: past history and current trends. *Medical Education*, *39*(4), 418–427. http://doi.org/10.1111/j.1365-2929.2005.02127.x
Norman, G. R., Monteiro, S. D., Sherbino, J., Ilgen, J. S., Schmidt, H. G., & Mamede, S. (2017). The causes of errors in clinical reasoning: cognitive biases, knowledge deficits, and dual process thinking. *Academic Medicine*, *92*(1), 23–30. http://doi.org/10.1097/ACM.0000000000001421
Nouh, T., Boutros, M., Gagnon, R., Reid, S., Leslie, K., Pace, D., … Meterissian, S. H. (2012). The script concordance test as a measure of clinical reasoning: A national validation study. *American Journal of Surgery*, *203*(4), 530–534. http://doi.org/10.1016/j.amjsurg.2011.11.006
Oren, C., Kennet-Cohen, T., Turvall, E., & Allalouf, A. (2014). Demonstrating the validity of three general scores of PET in predicting higher education achievement in Israel. *Psicothema*, *26*(1), 117–126. http://doi.org/10.7334/psicothema2013.257
Padilla, J.-L., & Benitez, I. (2014). Validity evidence based on response processes. *Psicothema*, *26*(1), 136–144. Retrieved from http://www.psicothema.com/pdf/4171.pdf
Page, G., Bordage, G., & Allen, T. (1995). Developing key-feature problems and examinations to assess clinical decision-making skills. *Academic Medicine : Journal of the Association of American Medical Colleges*, *70*(3), 194–201. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/7873006
Park, A. J., Barber, M. D., Bent, A. E., Dooley, Y. T., Dancz, C., Sutkin, G., & Jelovsek, J. E. (2010). Assessment of intraoperative judgment during gynecologic surgery using the Script Concordance Test. *American Journal of Obstetrics and Gynecology*, *203*(3), 240.e1-6. http://doi.org/10.1016/j.ajog.2010.04.010
Pelaccia, T., Tardif, J., Triby, E., & Charlin, B. (2011). An analysis of clinical reasoning through a recent and comprehensive approach: the dual-process theory. *Medical Education Online*, *16*(1), 1–9. http://doi.org/10.3402/meo.v16i0.5890

Petrella, R. J., & Davis, P. (2007). Improving management of musculoskeletal disorders in primary care: the Joint Adventures Program. *Clinical Rheumatology*, *26*(7), 1061–1066. http://doi.org/10.1007/s10067-006-0446-4

Petrucci, A. M., Nouh, T., Boutros, M., Gagnon, R., & Meterissian, S. H. (2013). Assessing clinical judgment using the Script Concordance test: The importance of using specialty-specific experts to develop the scoring key. *American Journal of Surgery*, *205*(2), 137–140. http://doi.org/10.1016/j.amjsurg.2012.09.002

Regenbogen, S. E., Greenberg, C. C., Studdert, D. M., Lipsitz, S. R., Zinner, M. J., & Gawande, A. A. (2007). Patterns of technical error among surgical malpractice claims: an analysis of strategies to prevent injury to surgical patients. *Annals of Surgery*, *246*(5), 705–11. http://doi.org/10.1097/SLA.0b013e31815865f8

Reyna, V. F. (2008). A theory of medical decision making and health: fuzzy trace theory. *Medical Decision Making : An International Journal of the Society for Medical Decision Making*, *28*(6), 850–65. http://doi.org/10.1177/0272989X08327066

Reyna, V. F., & Adam, M. B. (2003). Fuzzy-Trace Theory, Risk Communication, and Product Labeling in Sexually Transmitted Diseases. *Risk Analysis*, *23*(2). Retrieved from https://pdfs.semanticscholar.org/72c9/29d2e7854f094af40dbf1cea32c8a574e81b.pdf

Rios, J., & Wells, C. (2014). Validity evidence based on internal structure. *Psicothema*, *26*(1), 108–116. http://doi.org/10.7334/psicothema2013.260

Rogers, S. O., Gawande, A. A., Kwaan, M., Puopolo, A. L., Yoon, C., Brennan, T. A., & Studdert, D. M. (2006). Analysis of surgical errors in closed malpractice claims at 4 liability insurers. *Surgery*, *140*(1), 25–33. http://doi.org/10.1016/j.surg.2006.01.008

Royal College of Physicians and Surgeons of Canada. (2005). *CanMEDS 2005 Framework*. Retrieved from http://www.royalcollege.ca/portal/page/portal/rc/common/documents/canmeds/framework/the_7_canmeds_roles_e.pdf

Royal College of Physicians and Surgeons of Canada. (2017). Objectives of Training in the Specialty of General Surgery. Retrieved May 1, 2018, from http://www.royalcollege.ca/cs/groups/public/documents/document/ltaw/mtyx/~edisp/rcp-00161418.pdf

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, *18*(2), 119–144. http://doi.org/10.1007/BF00117714

Schank, R. C. (1982). *Dynamic memory: a theory of reminding and learning in computers and people*. Cambridge University Press.

Schank, R. C., & Abelson, R. P. (1977). *Scripts, Plans, Goals, and Understanding*. Hillsdale: Lawrence Erlbaum Associates. Retrieved from https://www.amazon.com/Scripts-Plans-Goals-Understanding-Intelligence/dp/0898591384

Schmidt, H., Norman, G., & Boshuizen, H. (1990). A cognitive perspective on medical expertise: theory and implication. *Academic Medicine*, *65*(10), 611–21.

Schmitt, N. (1995). Uses and abuses of coefficient alpha. *Psychological Assessment*, *8*(4), 350–353.

Schön, D. A. (1987). *Educating the reflective practitioner: toward a new design for teaching and learning in the professions*. Jossey-Bass.

Schubert, A., Tetzlaff, J. E., Tan, M., Ryckman, V., & Mashca, E. (1999). *Consistency, Inter-rater Reliability, and Validity of 441 Consecutive Mock Oral Examinations in Anesthesiology: Implications for Use as a Tool for Assessment of Residents*. *Anesthesiology*

(Vol. 91). [American Society of Anesthesiologists, etc.]. Retrieved from http://anesthesiology.pubs.asahq.org/Article.aspx?articleid=1946463

Sharma, S. (2004). Objective assessment of technical skills in surgery: assessment should include decision making. *British Medical Journal*, *328*(403). Retrieved from http://www.bmj.com.proxy.bib.uottawa.ca/content/328/7436/403.1

Sibert, L., Charlin, B., Corcos, J., Gagnon, R., Grise, P., & Vleuten, C. van der. (2002). Stability of clinical reasoning assessment results with the Script Concordance test across two different linguistic, cultural and learning environments. *Medical Teacher*, *24*(5), 522–527. http://doi.org/10.1080/0142159021000012599

Sidhu, R. S., Grober, E. D., Musselman, L. J., & Reznick, R. K. (2004). Assessing competency in surgery: Where to begin? *Surgery*, *135*(1), 6–20. http://doi.org/10.1016/S0039-6060(03)00154-5

Sireci, S. (1988). The Construct of Content Validity. *Social Indicators Research*, *45*(1), 83–117. http://doi.org/10.1023/A

Sireci, S., & Faulkner-Bond, M. (2014). Validity evidence based on test content. *Psicothema*, *26*(1), 100–107. http://doi.org/10.7334/psicothema2013.256

Sjoukje van den Broek, W. E., Marianne van Asperen, B. V, Eugène Custers, B., Gerlof Valk, B. D., Olle Th ten Cate, B. J., S van den Broek, W. E., … Valk, G. D. (2012). Effects of two different instructional formats on scores and reliability of a script concordance test. *Perspectives on Medical Education*, *1*, 119–128. http://doi.org/10.1007/s40037-012-0017-0

Stiegler, M. P., & Gaba, D. M. (2015). Decision-making and cognitive strategies. *Simulation in Healthcare*, *10*(3), 133–8. http://doi.org/10.1097/SIH.0000000000000093

Szatmary, P., Arora, S., & N, S. (2010). To operate or not to operate: a multi-method analysis of decision-making in emergency surgery. *American Journal of Surgery*, *200*(2), 298–304.

The Royal College of Physicians and Surgeons of Canada. (2018). CanMEDS Role: Medical Expert. Retrieved July 2, 2018, from http://www.royalcollege.ca/rcsite/canmeds/framework/canmeds-role-medical-expert-e

Tourangeau, R. (1984). Cognitive sciences and survey methods. In T. Javine, M. Straf, J. Tanur, & R. Tourangeau (Eds.), *Cognitive aspects of survey design: Building a bridge between disciplines* (pp. 70–100). Washington DC: National Academy Press.

Townsend, C. M., Beauchamp, R. D., Evers, B. M., & Mattox, K. L. (2016). *Sabiston textbook of surgery: the biological basis of modern surgical practice*. (C. Townsend, D. R. Beauchamp, M. B. Evers, & K. Mattox, Eds.) (20th ed.). Elsevier.

Way, L. W., Stewart, L., Gantert, W., Liu, K., Lee, C. M., Whang, K., & Hunter, J. G. (2003). Causes and Prevention of Laparoscopic Bile Duct Injuries. *Annals of Surgery*, *237*(4), 460–469. http://doi.org/10.1097/01.SLA.0000060680.92690.E9

Willis, G. B. (1999). *Cognitive interviewing: a "how to" guide*. North Carolina. Retrieved from http://www.chime.ucla.edu/publications/docs/cognitive interviewing guide.pdf

Willis, G. B. (2015). *Analysis of the Cognitive Interview*. New York: Oxford University Press.

Willson, S., & Miller, K. (2014). Data Collection. In K. Miller, B. Chepp, S. Willson, & J. L. Padilla (Eds.), *Cognitive interviewing methodology* (pp. 15–34). Wiley. Retrieved from https://www.wiley.com/en-us/Cognitive+Interviewing+Methodology-p-9781118383544

Wilson, A. B., Pike, G. R., & Humbert, A. J. (2014). Analyzing script concordance test scoring methods and items by difficulty and type. *Teaching and Learning in Medicine*, *26*(2), 135–145. http://doi.org/10.1080/10401334.2014.884464

Wood, T. J. (2014). In reply to McLaughlin. *Advances in Health Sciences Education*, *19*(3), 433–434. http://doi.org/10.1007/s10459-014-9520-x

Yates, J. F. (Jacques F. (2003). *Decision management: how to assure better decisions in your company*. Jossey-Bass.

Yates, J. F. (Jacques F., & Tschirnhart, M. D. (2006). Decision-Making. In K. Ericsson, N. Charners, P. Feltovich, & R. Hoffman (Eds.), *The Cambridge Handbook of Expertise and Expert Performance* (pp. 421–37). Cambridge: Cambridge University Press.

Yule, S., Flin, R., Paterson-Brown, S., Maran, N., & Rowley, D. (2006). Development of a rating system for surgeons' non-technical skills. *Medical Education*, *40*(11), 1098–1104. http://doi.org/10.1111/j.1365-2929.2006.02610.x

Appendices

Appendix 1. Delphi Instructions to Experts

Introduction

This study aims to explore assessment of clinical decision-making skills using the Script Concordance Test (SCT). The SCT is a written test that presents a clinical case and an accompanying series of questions that each present an investigative or management option followed by a new clinical finding, and asks examinees to use a Likert scale to indicate how the clinical finding influences the proposed investigative/management option. The current SCT consists of 43 clinical cases.

Instructions

As an expert on the Delphi panel you will be asked to select the most relevant clinical cases from the current SCT. Through an online survey using a 4-point Likert scale and comments, you will be asked to indicate the questions based on: 1) does the case actually address a realistic and challenging clinical scenario that incorporates uncertainty, 2) does it test decision-making skills, and 3) does it relate to a subspecialty of General Surgery defined by the Royal College of Surgeons of Canada.

All answers will be de-identified and your identity will not be known to the other participants on the panel. Feedback consisting of means and standard deviations for each item will be determined and items will be ranked ordered by their mean rating. This feedback will be sent back to you to rerate each item given this information. There will be a maximum of three rounds of rating to achieve consensus. Each round should take approximately 2 hours, for a total time commitment of 6 hours.

Appendix 2. Cognitive Interview Guide

Participant name:

Gender:

PGY level/staff:

1. Can you explain to me in your own words what this question is asking you/testing?
2. Was there any information you didn't know that you felt you needed to answer the question?
3. How did you arrive at the answer you chose?
4. What would have made you choose __ option?
5. Did you find this question difficult or easy? (If staff, also ask, do you think this question would be difficult or easy for a trainee? Why?)
6. Do you have any other comments about this question?

Appendix 3. Instructions to Staff and Resident Surgeons taking Final SCT

Introduction

This study aims to explore assessment of clinical decision-making skills using the Script Concordance Test (SCT). The SCT is a written test that presents a clinical case and an accompanying series of questions that each present an investigative or management option followed by a new clinical finding, and asks examinees to use a Likert scale to indicate how the clinical finding influences the proposed investigative/management option. All answers will be de-identified and your identity will not be known to the other participants. The test should take approximately 1-hour.