The Situational Judgment Test Validity Void: Describing Participant Response Processes

Michael David Lee Wolcott

A dissertation submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Education in the Learning Sciences and Psychological Studies program in the School of Education.

Chapel Hill
2018

Approved by:

Gregory J. Cizek

Thurston Domina

Robert Hubal

Jacqueline E. McLaughlin

Adam Meade

ABSTRACT

Michael David Lee Wolcott: The Situational Judgment Test Validity Void:
Describing Participant Response Processes
(Under the direction of Dr. Gregory J. Cizek)

Situational judgment tests (SJTs) are used to measure components of professional competence that cannot be assessed via traditional tests of knowledge and skill. Despite their popularity, there is a significant gap in the validity evidence and research on the response process to support how SJTs measure their intended constructs. This study evaluated an SJT to examine: (1) the factors that influence the response process, (2) the role of experience, (3) the role of contextual features, and (4) whether individuals attempt to identify the construct being assessed. Thirty participants—15 students and 15 pharmacists—completed a 12-item SJT designed to measure empathy. Each participant engaged in a think-aloud interview during the SJT followed by a cognitive interview that asked questions about their decision-making process. Results of the qualitative and quantitative analyses suggest that the SJT response processes include the complex integration of comprehension, retrieval, judgments, and response selections. In addition, job-specific knowledge and experiences comprised a significant portion of the retrieval process. Moreover, there was evidence that SJTs are highly contextual and that item characteristics such as setting, actors, or relationships can influence the response process. There was limited evidence to suggest individuals attempt to identify the construct being assessed. In summary, this study provides a comprehensive evaluation of the response process involved in SJTs and it contributes to foundational steps to generate validity evidence necessary to aid score interpretation.

To all those who said I could and—especially—to those who said I could not.

ACKNOWLEDGEMENTS

Thank you to all of those who made this dissertation possible. First, thank you to my dissertation advisor, Dr. Gregory Cizek, and committee members, Dr. Thurston Domina, Dr. Robert Hubal, Dr. Jacqueline McLaughlin, and Dr. Adam Meade. They each provided invaluable insight throughout this process. In addition, I want to thank my classmates and outstanding educators I have been fortunate to work with at the UNC School of Education; you all have been an inspiration and constant source of support in my growth as an educational researcher. Additional thanks to Dr. Jacqueline McLaughlin, who has served as an exceptional mentor and role model while working at the UNC Eshelman School of Pharmacy. Her support, along with assistance from Dr. Tom Angelo and Dr. Scott Singleton, aided my transition into the social sciences and I am grateful for the opportunity to work under their guidance. I also want to thank Provost Robert Blouin for launching this partnership between the Schools and for encouraging me to pursue this path. Dr. Mary Roth McClurg and Dean Dhiren Thakker also deserve special thanks for providing me with the opportunities to advance my training at the UNC Eshelman School of Pharmacy. Thank you to Dr. Jacqueline Zeeman for assisting with this research and to the faculty who helped design this SJT. Moreover, thank you to Nikki Lobczowski—you inspire me each day to do my best and we challenge one another to reach our fullest potential; you have been my guide throughout this process as a classmate, co-researcher, co-presenter, and best friend—I can never thank you enough. I also thank my parents, family, and phriends for constantly supporting my love of learning. And to Joel—for being there when it mattered most.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

Chapter 1: Introduction

Health Professions and the Non-Cognitive Dilemma

For decades health professions education has focused on the attainment of clinical knowledge as the dominant indicator of practitioner competence (Berwick & Finkelstein, 2010). This notion, however, has become less appealing due to the dramatic evolution of medicine over the years. Advances in technological capabilities as well as constant changes to our understanding of the human body have created an expansive field of knowledge that is difficult to understand, let alone master, during a student's time in school. The current drive in health professions educational reform is to prepare future healthcare providers with the knowledge, skills, and abilities to handle complex, ill-defined problems and situations encountered in practice (Cooke, Irby, & O'Brien, 2010; Irby, 2011).

A critical element of health professions educational reform is a paradigm shift that academic performance–the previous indicator of competence–is now a necessary, but not sufficient, quality of a competent clinician (Patterson et al., 2016). This shift has contributed to greater emphasis on the remaining aspects of performance, which are attributes of a separate entity known as professional competence. Professional competence is distinctly different than clinical competence, which refers to the knowledge and skills related to diagnosis, clinical decision making, and treatment management (Miller, 1990; Neufeld & Norman, 1985). Professional competence is a broad domain that is frequently described as the non-cognitive or non-academic qualities of practitioners that are necessary to optimize clinical care supporting

clinical knowledge, skills, and abilities (Epstein & Hundert, 2002; Farrington et al., 2012). The research presented here focuses on professional, rather than clinical, competence.

In health professions education, the term *non-cognitive* often refers to the interpersonal, intrapersonal, psychosocial, and behavioral knowledge, skills, and abilities necessary to effectively deliver healthcare. As stated, these are often broadly classified under an expansive domain labeled *professional competence* (Bardes, Best, Kremer, & Dienstag, 2009; Epstein & Hundert, 2002). Example qualities include a practitioner's motivation, integrity, empathy, confidence, and self-regulation. Understanding of and appreciation for how each of these qualities contributes to effective patient care is the minimum expectation of beginning practitioners; beyond this, practitioners should aspire to master knowledge, skills, and abilities related to professional competence throughout the course of their career (Levine & Cayea, 2015).

A review of the literature suggests there are variable conceptualizations of professional competence (Epstein & Hundert, 2002; Goldstein et al., 2006; Li, Ding, Zhang, Lie, & Wen, 2017). Of those discovered, the Professional Attributes Framework (see Table 1) offers a comprehensive outline of the proposed knowledge, skills, and abilities that comprise professional competence. Originally developed in the United Kingdom (UK), the framework is based on observational studies of first-year physicians. Importantly, it outlines the professional attributes physicians are expected to master and is applicable to the range of health professions. This framework is specifically used in selection procedures for the Foundation Programme—the UK's equivalent to graduate training in the United States (Patterson, Ashworth, Kerrin, & O'Neill, 2013). Standardized assessments for selection into the Foundation Programme, for example, target five of the eight sub-domains determined to be most salient: (1) patient focus (i.e. empathy), (2) coping with pressure (i.e. adaptability and prioritization), (3) working effectively

as a team, (4) commitment to professionalism, and (5) effective communication.

Table 1

Professional Attribute Framework (adapted from Patterson, Ashworth, Kerrin, & O'Neill, 2013)

Constructs/Sub-Domains	Definition	Example Behaviors/Scenarios
Patient focus (i.e. empathy)	Ensures patient is the focus of care. Demonstrates understanding and appreciation of the needs of all patients, showing respect at all times. Takes time to build relationships with patients, demonstrating courtesy, empathy and compassion. Works in partnership with patients about their care.	Identifying patient's views and concerns Considering patient needs outside of your own Empathizing with the patient
Coping with pressure (i.e. adaptability)	Capability to work under pressure and remain resilient. Demonstrates ability to adapt to changing circumstances and manage uncertainty. Remains calm when faced with confrontation. Develops and employs appropriate coping strategies and demonstrates judgement under pressure.	How to respond when you make a mistake Dealing with confrontation Seeking help
Working effectively as part of a team	Capability & willingness to work effectively in partnership with others and in multi-disciplinary teams. Demonstrates a facilitative, collaborative approach, respecting others' views. Offers support and advice, sharing tasks appropriately. Demonstrates an understanding of own and others' roles within the team and consults with others where appropriate.	Recognize and value other staff members Consult with colleagues about workflow and expectations Offer assistance to support colleagues
Commitment to professionalism	Displays honesty, integrity and awareness of confidentiality & ethical issues. Is trustworthy and reliable. Demonstrates commitment and enthusiasm for role. Willing to challenge unacceptable behavior or behavior that threatens patient safety, when appropriate. Takes responsibility for own actions.	Issues of confidentiality Challenging inappropriate behavior Commitment to learning
Effective communication	Actively and clearly engages patients and colleagues in equal/open dialogue. Demonstrates active listening. Communicates verbal and written information concisely and with clarity. Adapts style of communication according to individual needs and context. Able to negotiate with colleagues & patients effectively.	Gathering information and communicating intentions Negotiation skills Listening and communicating with different populations
Organization and planning*	Manages and plans workload effectively, displaying efficient time management and delivering tasks on time. Able to prioritize effectively and re-prioritize where appropriate. Is conscientious and maintains accurate records.	Effective time management Prioritize tasks effectively Maintains accurate records Manages plans and workload
Problem solving and decision making*	Demonstrates an ability to assimilate a range of information and identify key issues. Engages with the wider issues and thinks creatively to solve problems and reach appropriate decisions. Is proactive and demonstrates initiative. Is able to attend to detail.	Makes informed decisions Demonstrates initiative Assimilate and integrate information Attention to details
Self-awareness and insight**	Demonstrates awareness of the boundaries of their own competence and willing to seek help when required, recognizing that this is not a weakness. Exhibits appropriate level of confidence and accepts challenges to own knowledge.	Seek help when needed Admit a lack of knowledge Recognize boundaries of competence Accepts challenges Accepts mistakes
Learning and professional development***	Demonstrates desire and enthusiasm for continued learning, takes responsibility for own development. Willing to learn from others and from experience. Is open and accepting of feedback. Demonstrates a desire and willingness to teach others.	Enthusiasm to learn Learns from experience and mistake Accepts feedback

Notes: *Considered implicit to the situational judgment test methodology (not included as a construct)
**Considered to be integral to coping with pressure (subsequently consolidated)
***Considered to be integral to commitment to professionalism

Where students learn about these sub-domains is an emerging topic in the health professions: it was often assumed that developing professional competence was part of the hidden curriculum within the health professions (Hafferty, O'Donnell, & Baldwin Jr., 2015). More recently, the health professions are making these skills explicit and integrating them within their curricula to ensure they are acquired during their education and practice experiences (Goldstein et al., 2006).

Although assessment of professional competence is increasingly popular in the health professions, it remains a formidable challenge (Ferguson & Lievens, 2017; Patterson, Cleland, & Cousans, 2017; Ratanawongsa et al., 2006). Assessment of professional competence, although identified as relevant, often remains a secondary criterion for evaluation compared to the development of clinical knowledge, skills, and abilities (Kane, Clauser, & Kane, 2017). Moreover, assessment of professional competence is difficult because these sub-domains can overlap substantially, have variable definitions, and are not all considered equally important across the professions (Hays, 2013). Overall, the prioritization of other skill sets, inconsistency in the definitions, and mixed relationships among the sub-domains of interest has led to a fragmented field advancing in multiple directions.

Fortunately, describing and assessing professional competence in the health professions has become more focused largely in part to interests in improving admission processes at health professions schools and postgraduate training programs (Bardes, Best, Kremer, & Dienstag, 2009; Patterson, Ashworth, Kerrin, & O'Neill, 2013). As students become increasingly qualified for selection, differentiation among candidates becomes more critical (Patterson, Cleland, & Cousans, 2017). The assessment of professional competence, therefore, serves as an additional strategy to differentiate among candidates and assess their readiness for professional training

(Patterson, Cleland, & Cousans, 2017). The assessment approaches adopted and evaluated in the health professions fields have greatly improved the variety of instruments available to analyze sub-domains of professional competence (Li, Ding, Zhang, Liu, & Wen, 2017).

The importance of evaluating professional competence has also diffused beyond selection to describe growth throughout curricula, predict academic and practice performance, and identify areas for personal improvement (Cowart, Dell, Rodriguez-Snapp, & Petrelli, 2016; Goss et al., 2017; Persky, Greene, Anksorus, Fuller, & McLaughlin, 2017). The current challenge is distinguishing which assessment strategies offer valid data in describing participant professional competence while balancing administrative and feasibility limitations. Table 2 summarizes key characteristics of approaches commonly used to measure professional competence in health professions education in addition to their strengths and weaknesses.

Table 2

Common Approaches for Measuring Professional Competence

	Single Construct Questionnaires	Personality Assessments (Hojat, Erdmann, & Gonnella, 2013)	Multiple Mini Interview (Rees et al., 2016)	Situational Judgment Tests (Patterson, Zibarras, & Ashworth, 2015)
Content	Variable based on the instrument (e.g., Jefferson scale of empathy, emotional intelligence, etc.)	Questions oriented to quantify levels of personality traits (e.g., the NEO Big 5, HEXACO, etc.)	Scenarios are oriented to evaluate a construct of interest (e.g. adaptability, integrity, empathy)	Scenarios are designed to evaluate a construct of interest (e.g. adaptability, integrity, empathy)
Format	Questionnaire / survey completed by the individual	Questionnaire / survey completed by the individual	Individuals discuss their response to a scenario with a rater	Examinees select or rank optimal responses to a presented scenario
Scoring Process	Variable; points related to presence of a quality / attribute	Variable; points related to presence of a quality / attribute	Rater scores the individual based on observed discussion	Variable; points assigned based on consensus with a key
Advantages	- Often short / brief instruments - Focuses on a specific construct	- Substantial research - Generally stable construct - Numerous uses	- High validity & reliability - Can assess multiple constructs at once	- Moderate validity & reliability - Can assess multiple constructs at once
Disadvantages	- Potential for faking - Requires multiple surveys if various skills - Variable quality	- Often require commercial licenses - May oversimplify personality traits - Can be lengthy	- Resource intensive - Potential for faking - Influence of rater bias	- Potential for faking - Highly variable design and scoring strategies

Situational Judgment Tests

Of the assessment strategies described in Table 2, situational judgment tests (SJT) are a recent addition that has attracted substantial interest. SJTs originated in personnel selection to evaluate skills beyond cognitive ability, which was previously used to predict occupational performance (Campion, Ployhart, & MacKenzie Jr., 2014; Chan & Schmitt, 2002). SJTs were first characterized as a low-fidelity simulation (Motowidlo, Dunnette, & Carter, 1990) intended to measure how potential employees would respond to scenarios encountered on the job. Motowidlo and colleagues (1990) contributed to the initial evidence supporting that SJTs could measure unique skill sets different from cognitive ability–which could inform hiring decisions.

Growing interest in SJTs is due to its similarity with multiple mini-interviews (MMIs), which are used extensively in health professions education and selection (Patterson et al., 2016). During both an SJT and MMI, a participant is presented with a scenario commonly-encountered in practice and is requested to describe how to respond. A key difference between MMIs and SJTs is that an MMI includes an interaction that is evaluated by an interviewer, whereas an SJT is administered electronically or as a paper-and-pencil test. MMIs, therefore, are sometimes viewed unfavorably because they involve subjective scoring techniques, are highly resource intensive, and are administratively complex (Rees et al., 2016). As a result, SJTs are being investigated as a complementary assessment methodology for large-scale testing of professional competence, especially in the setting of graduate and postgraduate admissions (Koczwara et al., 2012; Patterson et al., 2016).

SJTs are designed to evaluate how an examinee would respond to situations commonly encountered in practice. During an SJT, the examinee is presented with a hypothetical scenario, which may be based on job analyses, critical incidents, or personal experiences of the test

developers (Weekley & Ployhart, 2006) along with a question (i.e., test item) about the scenario and multiple response options that represent potential actions. Each action is evaluated by the examinee, who is asked to address the likelihood they should perform the action or who is asked to evaluate the effectiveness of the action. Items are designed to capture the knowledge of the examinee in selecting the most appropriate response options that would be consistent with the expectations of the job, which are tied to the constructs being measured. The constructs of interest targeted by an SJT will vary based on the job (e.g. management, healthcare, etc.); the common goal, however, is to measure participant attributes such as professional competence that are different from qualities such as cognitive ability.

At the end of an SJT, participants are assigned a score based on how well their response selections align with a key. The key used to score performance on an SJT can be developed by aggregating response data from subject matter experts (e.g. experienced clinicians), known as a *rational key*, or from the most common examinee responses, known as an *empirical key*. The use of a rational key is the prominent scoring method in SJT research (De Leng et al., 2017). High scores indicate a participant has high levels of the trait being evaluated (i.e. knowledge pertaining to the construct or constructs of interest), which is inferred by how close examinee performance matches that of a job expert when they approach the same tasks.

Debate Surrounding the Situational Judgment Test

Despite its simplicity, debate surrounding SJTs has been labeled a “hot mess” due to the rapidly changing foci about what matters in the field of SJT research as well as the evolving theoretical and empirical support of SJTs (McDaniel, List, & Kepes, 2016, p. 47). Discourse about SJTs suggests a lack of consensus on the salient design features to best measure the constructs of interest or a mutual misunderstanding of the response process when participants

complete an SJT. This section outlines reasons for this debate, the predominant opinions, and the impacts on current SJT research.

A hallmark of this debate is understanding the role of instrument design strategies in ensuring SJTs produce high-quality data. Difficulty reaching consensus among researchers is due considerably to the versatility of SJTs as instruments, which is also one of the benefits of using SJTs. For instance, users can tailor an SJT design process to fit their needs based on test developer preferences or context-specific requirements. A consequence of this, however, is that comparison across SJTs is often problematic and it is difficult to ensure SJT design processes meet quality standards (McDaniel, Hartman, Whetzel, & Grubb, 2007; Patterson, Zibarras, & Ashworth, 2016). Currently, researchers are encouraged to maximize reliability and validity of the data by incorporating design principles supported by evidence (Lievens & Patterson, 2011; Lievens, Peeters, & Schollaert, 2008; McDaniel, Morgenson, Finnegan, Campion, & Braverman, 2001; McDaniel & Nguyen, 2001); these principles are described in greater detail in Chapter 3 to inform SJT design for this research.

SJT design is critical because it influences the interpretation of individual performance and empirical findings. Originally, SJTs were presumed to produce quality measures of the constructs of interest because the scenarios were generated from on-the-job reports or experts agreed the situations were consistent with practice (McDaniel, Whetzel, Hartman, Nguyen, & Grubb, 2006). Upon further psychometric analyses, however, SJTs came to be seen as multidimensional instruments measuring complex skill sets that were difficult to reliably separate due to significant overlap and poor definitions of the constructs (Sorrel et al., 2016). Consequently, it was difficult to know if an SJT was truly measuring the constructs of interest or

if performance was an artifact of attributes unbeknownst to the researcher, such as an interaction of multiple constructs or poorly designed test items.

Lievens (2017) argues that SJTs can be designed purposefully to measure a construct of interest: an approach he calls *construct-driven SJTs*. This approach to SJT design incorporates scenarios generated from practice experiences while also integrating theoretical and empirical understanding of the sub-domains of professional competence. In other words, scenarios and response options are crafted to tap into salient features of the construct of interest based on evidence in the literature of what components of the construct should be present if we are truly measuring that construct. This approach is intended to create SJT items that are unidimensional and more consistent with theoretical underpinnings—a consideration that was frequently ignored in prior SJT research.

Moreover, the construct-driven approach to SJT development parallels the *evidence-centered design* approach that was founded in educational assessment (Mislevy, Almond, & Lukas, 2003; Riconscente, Mislevy, & Corrigan, 2016). Both methodologies stress the importance of a systematic approach in defining the construct to be tested, outlining the components, and ensuring alignment between test items and the construct of interest. The research described here utilized the construct-driven approach to develop scenarios and response options consistent with theoretical elements of the construct of interest to ensure an SJT measures the intended construct; the extent to which there is measurement fidelity to the construct of interest was evaluated through investigation of participants' response processes, outlined in subsequent chapters.

Discourse about SJTs also centers on understanding the role of knowledge, experiences, and other antecedents in the response process when participants complete an SJT. Again,

diversity in SJT methodology has led to difficultly in forming a theoretical framework that elucidates what contributes to SJT performance. As previously described, an SJT is designed to have participants reflect on their knowledge and experiences to inform their decision-making processes as they select an action intended to produce a desirable outcome. Currently, this response process is believed to be influenced by participant ability, interests, personality, values, emotional intelligence, and job-specific as well as general knowledge and experience (Lievens & Motowidlo, 2016).

The extent to which these factors influence SJT performance is highly dependent on design features, which is described in greater detail in the following chapters. The setting or contextual information provided in an SJT item, for example, has been an area of recent focus. Lievens and Motowidlo (2016) argue SJTs may not be as situational as previously suspected. They argue the item setting can be stripped from an SJT item and not dramatically influence examinee performance, which implies the setting may not be critical. This argument, however, had not been sufficiently explored. It is plausible SJT questions may not draw on job-specific knowledge or the setting may not influence the recall or decision-making process engaged to select an appropriate course of action.

Moreover, research has begun to explore the examinee's *ability to identify criteria* (ATIC) when they respond to instruments like an SJT (Griffin, 2014; Kleinmann et al., 2011). ATIC research suggests that a crucial element in the response process is whether the candidate can identify the attribute that is being evaluated by an item or task. For example, a candidate reviews an SJT item and speculates it is measuring a construct such as adaptability or empathy based on the presented information. Recognition of the construct then informs their response option to address the task by correctly matching the response based on the need in the scenario.

Overall, the research presented here explored how these attributes (i.e. job-specific knowledge, item setting, and the ability to identify the construct) could influence the response process and provided additional evidence regarding their roles in SJT performance.

The Precarious Position of Situational Judgment Tests

Escalating interest in SJTs initially eclipsed efforts to generate supporting evidence for its use as an assessment strategy in the health professions. Gessner and Klimoski (2006) noted the interest in and potential of SJTs as useful instruments overshadowed the necessary investigations describing theoretical underpinnings of SJTs. Moreover, there were inadequate attempts to establish validity evidence that distinguished what constructs were assessed and the elements involved in response processes; given the enthusiasm SJTs are now heavily used without sufficient evidence to support its systemic use (Sorrel et al., 2016). Overall, questions remain about the quality of data an SJT produces and a deeper understanding of what is being measured (e.g. the construct of interest and associated cognitive processes) is needed.

Although these concerns may seem trivial to those outside the measurement specialty, SJT use can be consequential. SJTs are being applied more often in high-stakes environments such as selection into health professions education and postgraduate training; the mismeasurement of attributes in these arenas may have serious consequences for examinees in addition to the wellbeing of others. It is imperative that assessments informing high-stakes decisions have sufficient validity evidence to support their interpretation and use (Caines, Bridglall, & Chatterji, 2014).

Most validity evidence supporting SJT score interpretations is generated using quantitative methodologies focused on correlations with other variables and measures of internal structure (Christian, Edwards, & Bradley, 2010; Weekley & Ployhart, 2005). The nearly

exclusive emphasis on quantitative validity evidence, however, limits a comprehensive understanding of the cognitive processes involved when completing SJT items. Considering an SJT is intended to measure a decision-making or judgmental process, there are various processes that are assumed to take place that have not been thoroughly described in the literature. A void exists in the literature regarding SJT response processes that could be explored further with appropriate methodologies.

SJTs incorporate complex, contextualized situations and a host of responses that can vary substantially based on the constructs of interest being assessed. Emerging research on assessing similar, complex skills demonstrates that alignment and design of instruments to evaluate these knowledge, skills, and abilities is challenging but not insurmountable (Erickan & Oliveri, 2016). It is possible to accomplish measurement of these attributes by outlining the intricacies and connections of constructs to be assessed, the processes individuals are expected to engage, and the meaning of the findings (Geisinger, 2016). Care and colleagues (2016) recommend deconstructing how individuals approach problems from a cognitive and social perspective, which is paramount as it applies to SJTs. A greater understanding of the response processes and attributes that influence those processes addresses a component of the many challenges associated with SJTs; however, it can also greatly inform SJT design and research to ensure it yields valid inferences about the intended constructs being assessed.

Purpose and Research Questions

There remains a void in the theoretical and empirical understanding of the response processes involved when completing an SJT that targets a specific construct. Few studies have examined the cognitive processes involved when examinees take an SJT. The *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA],

American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014) identify *evidence based on response processes* as a crucial element of validity evidence for assessments that require complex thinking or decision-making. Validity evidence to support a claim that the assessment measures the intended construct and to confirm examinees interpret the assessment appropriately is critical but has been neglected in SJT research. Additional research is needed to explore the response process examinees engage in during an SJT to thoroughly describe what elements of knowledge, skills, and abilities are activated in this process.

The purpose of this study was to address a gap in the validity evidence supporting SJTs in assessing constructs of interest. This study focused on generating evidence of response processes to an SJT measuring one construct that is typically of interest in health professions training: empathy. This research provides a prototype for exploring and describing response processes when using construct-driven SJTs with the intent of applying the methodology to SJTs measuring other constructs of interest in the future. The research questions included:

RQ1: What factors and strategies are involved in the cognitive processes when examinees respond to SJT items?

RQ2: What is the role of job-specific experiences (i.e. student or experienced clinicians) in the response process to SJT items?

RQ3: What is the role of the setting presented in SJT items in the response process (i.e. the influence of healthcare specific setting or non-healthcare specific setting)?

RQ4: What is the role of the ability to identify the construct being evaluated (i.e. empathy) in the response process to SJT items?

Due to the limited evidence regarding the response processes when completing an SJT, the research questions were exploratory in nature and there were minimal hypotheses about the findings. With regards to the first research question, it was anticipated that components pertaining to the theoretical underpinnings of SJTs (e.g. values, interest, ability, prosocial behaviors, etc.) may be evident in the response process of examinees. Although this may not be made explicit by participants, the goal was to probe participants to better understand how they believed these factors may contribute. For the second research question, it was suspected that greater job-specific experiences would influence response processes in that individuals would recall these experiences more often to address the presented situations and pick an optimal response that draws more heavily from those experiences. The third research question was intended to describe whether the setting presented in the item was able to influence response process. Prior research would suggest the setting is not a critical element and this work was expected to clarify this further as it pertains to response processes (Krumm et al., 2015; Rockstuhl et al., 2014). The final research question was to initiate an understanding of whether participants can identify the construct of interest being measured and the potential influence on response processes. It was suspected individuals would include this aspect in their response, although it was unclear if this would be at the forefront of or influences their thought processes.

Summary

The use of SJTs in the health professions is a rapidly growing phenomenon as an approach to measure professional competence, which is difficult to capture and often resource intensive to measure using other methodologies. Despite its popularity as a tool and expansive research in industrial and organizational psychology, there remain deficits in the validity evidence supporting how SJTs measure constructs of interest with regard to the response process.

SJT research is scarce regarding the response processes examinees use to respond to the scenarios they are presented, which can have profound implications on future use in practice. The goal of this research was to explore the response processes used during an SJT and outline what factors may influence these processes.

Chapter 2: Literature Review

Four areas of research inform the background and design of this research study. First, an overview of the theoretical basis of SJTs is provided to describe elements hypothesized or demonstrated to influence examinee performance. Second, a review of pertinent validity evidence supporting SJTs is presented to identify gaps with an emphasis on what is to be addressed through this research. Third, a summary of research on complex response processes used during assessments will outline the potential response processes examinees may engage with when completing an SJT.

Fourth, the chapter concludes with a brief review of the theoretical and empirical understanding of the construct evaluated in this study: empathy. Empathy was selected as the construct of interest for this research because there is substantial evidence that empathy expressed by health care professionals enhances patient satisfaction, comfort, and trust, which can contribute to positive patient outcomes (Kim, Kaplowitz, & Johnson, 2004; Reiss et al., 2008). Moreover, Quince and colleagues (2016) suggest that empathy is becoming as important in healthcare as clinical competence.

In summary, the chapter presents several models that describe SJTs, response processes, and empathy. Each of these models will inform a combination of study design elements: SJT design, interview protocols, and coding schemes used for qualitative data analysis.

Theoretical Basis of the Situational Judgment Test

The resurgence of SJTs has led to a greater focus on its theoretical underpinnings (Lievens & Motowidlo, 2016). In general, there are a host of antecedents that influence SJT performance regardless of the construct being evaluated or the context of the test. The theoretical model has been refined over the years with the most prominent being a model crafted by Lievens and Motowidlo (2016). In this model, shown in Figure 1, they identify attributes such as emotional intelligence, interests, values, personality traits, cognitive ability, and experiences (both general and specific to the job) as critical precursors to informing decision making processes on an SJT. Recognizing these antecedents was significant to this research as the aim was to describe the response process during an SJT to determine how these antecedents may influence the response process and subsequently be described during participant interviews. The antecedents are assumed to or have been shown to relate to SJT performance; therefore, these elements will be included in the coding schemes described in Chapter 3 and inform qualitative data analysis.

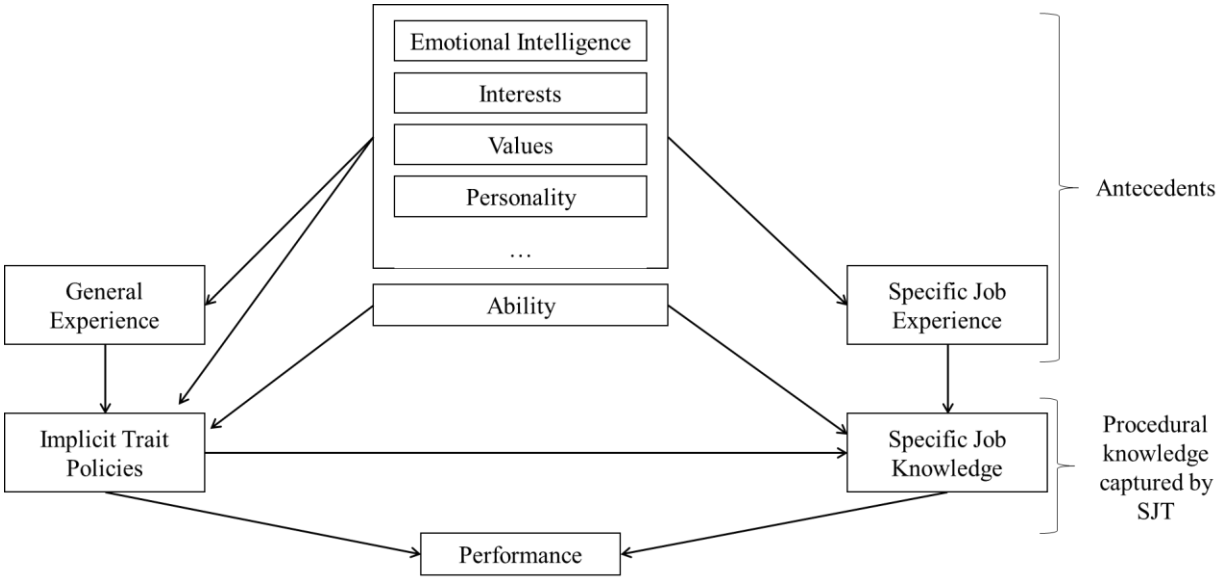


Figure 1. Model of knowledge determinants and antecedents of situational judgment tests (adopted from Lievens & Motowidlo, 2016)

The consensus has been that SJTs measure procedural knowledge regarding effective actions in response to scenarios presented on a test (Clevenger, Pereira, Wiechmann, Schmitt, & Harvey, 2001; McDaniel & Nguyen, 2001). Procedural knowledge in this setting refers to knowledge about how to respond to a scenario not necessarily whether individuals possess an ability to carry out the response in person. In addition, procedural knowledge could include when or how to apply that knowledge based on the presented scenario.

Motowidlo and Beier (2010) advanced this theory based on the understanding of knowledge acquisition (Beier & Ackerman, 2005; Hambrick, 2003) to suggest that procedural knowledge includes two types: general domain knowledge and specific job knowledge. *General domain knowledge* refers to the appreciation of the costs and benefits of expressing a trait (i.e. following a certain action) in response to a scenario. General domain knowledge reflects the fundamental socialization processes and personal dispositions that are not obtained through job-specific experiences. *Specific job knowledge* can be learned only through that particular job or jobs like it (Lievens & Motowidlo, 2016). With respect to the health professions, this could include other service-oriented jobs such as human resources, teaching, social services, or public safety. Each of these knowledge antecedents plays a role in SJT and job performance; a study by Motowidlo and Beier (2010) showed both components predict job performance equally well.

Another important theoretical element is the relationship of general domain knowledge to *implicit trait policies* (ITPs), a concept introduced by Motowidlo, Hooper, and Jackson (2006). ITPs refer to the policies individuals use when weighing sources of information to make evaluative judgments. ITP theory suggests individuals will express certain traits depending on the situation and the perceived cost and benefits associated with their behaviors based on their general domain knowledge. The decision to express a trait (i.e. chose an action that seems most

appropriate) is often mediated by other characteristics of the individual such as personality, values, interests, and experiences (Motowidlo, Hooper, & Jackson, 2006).

During an SJT, people who have higher tendencies for agreeableness may gravitate towards response options that appear to be more agreeable. This could introduce construct-irrelevant variance if agreeableness was not considered to be a component of the construct of interest. The relationships of ITPs to general domain knowledge and SJT performance has been essential in understanding observed relationships between personality and SJT performance. ITPs are presumed to vary across individuals, similar to personality traits.

A series of three studies by Motowidlo, Hooper, and Jackson (2006) confirmed there was a positive albeit weak relationship between ITPs measured by SJTs to certain personality attributes. Their series of three studies included the development of an SJT for managers that had response options tailored to target extraversion, agreeableness, and conscientiousness. Their hypothesis was that participants would identify a response option as more effective if it was consistent with their personality traits; in other words, a person who is more extraverted would be more likely to rate a response option as highly effective if that response option was related to or expressed elements of extraversion. Two of the three studies included 196 undergraduates and showed the average correlation between ITPs and the associated personality traits (measured using the NEO Five-Factor Inventory) was .31 for agreeableness and .37 for extraversion; however, there was no significant relationship with conscientiousness.

The third study by Motowidlo and colleagues (2006) investigated the relationship of ITPs and participant behaviors in simulated work situations. Ninety-nine undergraduate students completed a simulation in which they addressed a concern of an actor who portrayed a coworker, subordinate, supervisor, or customer. The simulated interactions were rated by four research

assistants on the level of agreeableness and extraversion the participant displayed in their response. Individual differences in ITPs for agreeableness predicted agreeable behaviors with an average correlation of .33; the findings of the three studies suggested that ITPs can be related to personality traits and be expressed to varying degrees during work-related simulations and SJTs.

Another theoretical element is the extent to which SJTs are truly situational. The findings that general domain knowledge and specific job knowledge relate equally to SJT performance suggest that SJTs could be considered tests of general domain knowledge instead of situationally specific knowledge (Lievens & Motowidlo, 2016).

A collection of studies by Krumm and colleagues (2015) evaluated the effect of situational stems on SJT performance and showed inclusion of the situation descriptions may not be necessary for a majority of SJT items. In their first study, 436 participants (students and working people) were given a 35-item SJT intended to measure knowledge, skills, and abilities related to teamwork, such as conflict resolution, collaborative problem solving, communication, and goal setting. SJT items were modified to have a version with and without elaborate situation descriptions. The performance on the respective SJT items was compared and determined that the situation descriptions were not necessary for approximately 71% of the items.

The second study by Krumm (2015) investigated whether the effect was due to the content domain being tested (i.e. teamwork). This study included 557 pilots who completed a 30-item SJTs that had an equal number of questions measuring teamwork, employee integrity, and decision-making in flight scenarios (i.e. job-specific knowledge and skills). Across the three tests, it was determined that it did not make a significant difference in performance if situation descriptions were included for 63% of the items. Of note, there was a trend in the data that the specific construct being evaluated may have a role in whether the providing the situational

descriptor has a significant role. For instance, thirty percent of the items assessing job-specific knowledge and skills of pilots could not be answered without the situational descriptors. The researchers argue this may be related to context-specific courses of actions; in other words, certain actions may be warranted based on specific contextual cues that have to be provided in the situational descriptors. Therefore, the setting may play a role for certain constructs.

Overall, the findings suggest that general domain knowledge is sufficient to solve a majority of SJT items and the label SJT may be a misnomer (Lievens & Motowidlo, 2016). That conclusion is still highly debated, however (Fan, Stuhlman, Chen, & Weng, 2016; Harris, Siedor, Fan, Listyg, & Carter, 2016; Harvey, 2016; Melchers & Kleinmann, 2016). The present research is positioned to explore that question by describing the extent to which general experiences are retrieved in SJT response processes compared to job specific experiences.

Personnel selection research has recently described a new element that may play a critical role in the theoretical understanding of assessments intended to evaluate decision-making processes of candidates: the *ability to identify criteria* (ATIC). ATIC refers to a candidate's capacity to distinguish which construct is being evaluated in these types of scenarios (Griffin, 2014; Kleinmann et al., 2011). ATIC is believed to mediate participant responses by serving as a filter to guide their selections based on their cognitive ability, social understanding, and preparation for the testing, as shown in Figure 2.

A study on medical student selection by Griffin (2014) showed ATIC was predictive of student performance and that ATIC is an attribute that needs further exploration. Her study included 319 applicants for medical school at an Australian university. As part of the selection process for the medical school, students were required to participate in a multiple mini interview (MMI). During this assessment, candidates rotated through 9 interview stations; at each station

the candidate interacted one-on-one with an interviewer (i.e. actor) in a simulated scenario designed to measure constructs such as empathy, integrity, and adaptability. Candidates were scored by the interviewer on a 7-point scale with 1 indicating poor overall performance on the station and 7 representing an outstanding performance. At the end of each interview station candidates were asked to write down the main quality the interviewer was assessing. The candidate answers to this question were rated by two judges on a scale of 0 (low fit) to 3 (high fit) with the construct that was measured according to the MMI development committee. There was a significant weak positive correlation (.33) between ATIC and MMI scores; overall, the findings suggest that ATIC may be an influential component in selection assessment performance.

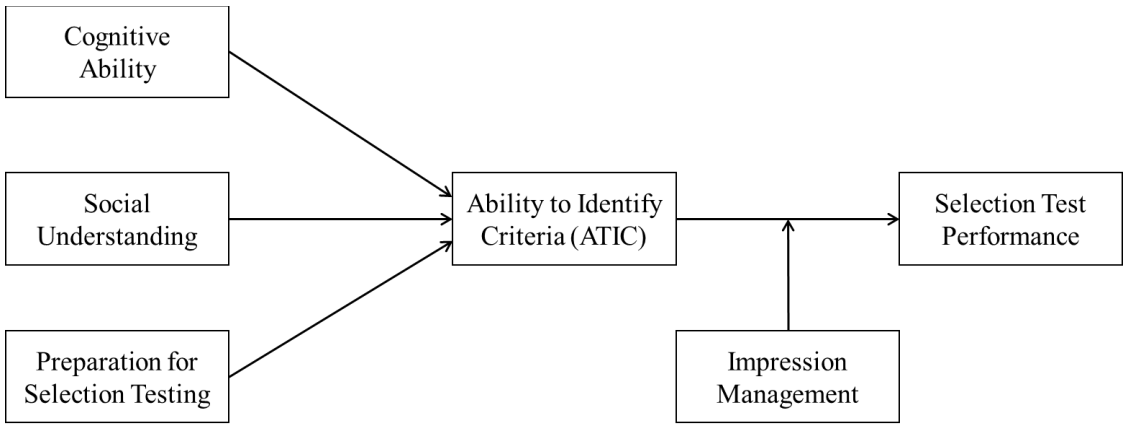


Figure 2. Model of ability to identify criteria in selection tests (adopted from Griffin, 2014)

The model presented by Griffin (2014) also notes the significance of *impression management*, which is extent to which the candidate modifies his or her response based on what is expected from the employer or the one administering the assessment. In high-stakes settings, impression management can play a significant role in how examinees select responses to ensure they meet the qualities sought by the tester (Bourdage, Wiltshire, Lee, & 2015; Cheng, Chiu, Chang, & Johnstone, 2014). In an SJT, for example, the examinee may select response options

that seem consistent with the mission and vision of the job setting or those that seem like an option the supervisor would choose instead of what he or she feels is best. Impression management is congruent with faking effects discussed later in this chapter.

Thus far, SJT researchers have yet to investigate how the ATIC contributes to individual performance, if at all. During an SJT, for example, it is likely that a person reads the scenario and thinks of plausible options based on their interpretation of what the question is asking them to do (e.g. empathize by staying late after work, adapt by responding to an emergent need, etc.). If the examinee is unable to discriminate between these constructs, he or she is less likely to respond to the question correctly; therefore, this can be a critical element in the response process. In the research proposed here, participants will be asked directly about what construct they believe is being assessed by each item and how that influences their decision-making process. It is also unknown if participants explicitly identify these constructs on their own or if recognition only emerges through specific probing. In addition, there may be differences in whether identification of the construct is more inherent with expert clinicians than with novices. This research aims to begin the exploration of these questions.

Validity Evidence for Situational Judgment Tests

Validation of any instrument involves the collection, synthesis, and evaluation of evidence gathered to support an intended interpretation of scores (Kane, 2016). Originally proposed by Messick (1989) and further refined by Kane (1992, 2006, 2013), the argument-based approach to validity specifies the necessary and sufficient conditions for validity using a structured framework. The *Standards* (AERA, APA, & NCME, 2014) specify five general sources of validity evidence: (1) content, (2) internal structure, (3) relationships with other constructs and criteria, (4) consequences of testing, and (5) cognitive / response processes.

These different sources of evidence contribute to the interpretation and use argument (IUA), which outlines the inferences connecting observed performance on the test to the proposed interpretations and use of the test scores (Kane, 2013). In other words, the IUA describes how test scores relate to the degree of mastery of the knowledge, skills, and abilities in the targeted domain. Of note, the focus of this research is to generate validity evidence in support of test score interpretation; therefore, there will be minimal emphasis on score use. There exists a smattering of validity evidence across these five sources of validity evidence that has been informative in supporting SJT score interpretation; however, these data are often fragmented and highly variable based on SJT design and context as discussed in the previous chapter (Christian, Edwards, & Bradley, 2010; Weekley & Ployhart, 2006). The following sections outline pertinent validity evidence relevant to SJTs according to the five sources suggested by the *Standards*. The review is not exhaustive; it focuses on major gaps in the validity evidence that can be addressed by this research.

Test content. *Standard* 1.11 (AERA, APA, & NCME, 2014) concerns validity evidence that supports the alignment of test content with the domain being tested. Standard design practice for SJTs begins with a definition of the domain to be assessed, which is usually a mixture of knowledge, skills, and abilities believed to be related to job performance. SJT items can be classified into four categories based on what constructs are being measured: (1) knowledge and skills, (2) applied social skills (e.g. leadership), (3) basic personality tendencies (e.g. empathy, integrity, etc.), or (4) heterogeneous composites (Christian, Edwards, & Bradley, 2010). SJT items on one test can include a mixture of these categories, which can create inconsistency in what is being measured. The goal of this research is to focus the test content on one construct of interest instead of a host of these categories.

Subject matter experts are frequently involved in the development of SJT scenarios, potential response options, and scoring keys based on their experiences (Christian, Edwards, & Bradley, 2010). Additional resources from employees can also be used to inspire test developers to create content relevant to the field such as job analyses, task inventories, and critical incident reports (Campion, Ployhart, & MacKenzie Jr., 2014). In general, this approach to SJT development has provided consistent evidence that the test content is highly related to the aspects of job performance considered relevant to the subject matter experts and applicants (McDaniel, Whetzel, Hartman, Nguyen, & Grubb, 2006).

Applicant reactions to SJTs are often positive and participants feel the scenarios reflect attributes of the job they are likely to experience (Bauer & Truxillo, 2006; Truxillo, Bauer, Campion, & Paronto, 2002). These reactions are attributed to the use of real-life example scenarios as well as the presentation mode of some SJTs (e.g. videos) that add a sense of realism and a sense of content alignment with job expectations. The contextual elements (i.e. setting) in SJT items are important for supporting test content validity evidence; however, it was previously highlighted there is debate whether SJTs need to be heavily situational. Lievens and Motowidlo (2016) argue that SJTs are capable of measuring job-specific and general domain knowledge independent of the setting. The present research will explore how setting influences the response process and the knowledge retrieved when taking an SJT.

Another element of the design process can limit validity evidence for SJTs—an SJT developed extensively based on subject matter expert knowledge and experience can neglect theoretical or empirical research that describes appropriate behaviors or strategies to respond to difficult scenarios in educational or workplace environments (Lievens, 2017). Relying on

experienced practitioners to define the appropriate response options to a scenario does not imply those are the best responses from a psychological, emotional, or social perspective.

It is recommended researchers develop SJTs that are cognizant of our theoretical understanding of social and behavioral constructs to ensure the elicited responses are consistent with evidence of what should be measured when evaluating those constructs. Lievens (2017) describes the construct-driven SJT as an approach to ensure greater alignment of the test with the construct being evaluated. He argues the items should be checked for their level of agreement with the theoretical understanding of the constructs (e.g. factor structures, definitions of the constructs, consistency with other measurements used to evaluate that construct). The presented research incorporated the construct-driven SJT design approach to ensure an SJT was measuring the appropriate elements of the tested constructed.

Internal structure. *Standards* 1.13, 1.14, and 1.15 (AERA, APA, & NCME, 2014) describe sources of validity evidence regarding how subsections or components of a test are unique as well as related to one another. Historically, the consensus has been that SJTs were often intentionally multidimensional because they measured a variety of constructs simultaneously that were highly related and difficult to distinguish from one another (Lievens, Peeters, & Schollaert, 2008). Psychometric techniques based on classical test theory have typically been used to analyze SJT performance data and the results have often been unremarkable. As might be expected, there was often evidence of multiple factors, but the structure could vary based on design principles, the context, or the constructs being assessed (Christian, Edwards, & Bradley, 2010).

An early systematic review conducted by McDaniel and colleagues (2001) collected internal reliability coefficients from all studies published on SJTs prior to 2000. The researchers

identified 33 coefficients that ranged from .43 to .94. This summary, however, did not control for the number of SJT items or response instructions; the data simply summarized the coefficients as part of a larger meta-analysis with minimal interpretation. As a result, Catano and colleagues (2012) conducted a more extensive meta-analysis. Their review identified 39 published studies from 1990 to 2011; these studies included a total of 45,062 SJT responses and 56 reliability coefficients. The studies included SJTs that ranged from 3 to 60 items in length and did not have a consistent type of response instruction. The meta-analysis corrected for sampling error to account for sample sizes and the weighted mean corrected *r* was .46. Overall, these findings show the internal consistency coefficients for SJTs were weak, especially considering their use in high-stakes decisions.

Again, the weak validity evidence on the internal structure of SJTs is attributable to the traditional design approach. Without a clearly defined domain, the design of SJTs could target various constructs that rarely minimized content overlap leading to poor internal consistency and complex factor structures. As described previously, construct-driven SJTs can address this deficiency in the validity evidence much how the evidence-centered design approach has been instrumental in improving educational assessments (Riconscente, Mislevy, & Corrigan, 2016). A construct-driven SJT is focused on creating items that are unidimensional because they target a specific construct instead of large domains of knowledge as was done previously (Guenole, Chernyshenko, & Weekly, 2017). Insights from psychologists as well as theoretical and empirical evidence to guide item design to target constructs of interest is posited to improve the validity evidence regarding internal structure by supporting more informed instrument design. The research proposed here will continue to build on this work by creating a construct-driven SJT focused on assessing empathy.

Relationships with Other Variables. *Standards* 1.16 through 1.24 (AERA, APA, & NCME, 2014) highlight the source of validity evidence involving the relationship of performance on SJTs relative to performance on other instruments or criteria for evaluation. SJT research in this area has been extensive in terms of describing what is measured. In the context of personnel selection, SJTs are often compared to job performance criteria to determine the incremental validity of SJTs versus other traditional measures in employee selection (e.g. interviews, assessment centers, etc.). SJTs have consistently shown to provide incremental validity above and beyond cognitive ability and personality measures in selection settings (Clevenger, Pereira, Wiechmann, Schmitt, & Schmidt-Harvey, 2001; Weekly & Ployhart, 2005).

The correlation of SJT performance with other attributes has been extensively studied but has yielded mixed results. SJT performance data is often evaluated for the degree of correlation with measures of cognitive ability, Big Five personality assessments, questionnaires measuring other constructs of interest, as well as rater assessments of performance on the job or in simulated scenarios (Guenole, Chernyshenko, & Weekly, 2017). In general, correlations of these measures with SJT tend to be relatively low but often statistically significant. The correlations, however, can vary substantially among studies based on the context, design, and constructs assessed by an SJT.

McDaniel and colleagues (2001), for example, showed SJT performance had a moderately positive relationship with cognitive ability (r = .46); however, they also noticed this could vary based on how the scenarios were generated (e.g. from a job analysis compared to critical incidents). Clevenger and colleagues (2001) suspect that variability in relationships to cognitive ability are reflective of the design processes and the situations presented, so these relationships should be interpreted with caution. In addition, a meta-analysis of the relationship

of SJT performance to personality traits showed agreeableness (r = *.25*), conscientiousness (r = .26), and emotional stability (r = .31) to have low positive correlations with performance (McDaniel & Nguyen, 2001).

Evidence supporting the relationship between SJT performance and other constructs has been problematic. A content analysis conducted by Christian and colleagues (2010) reported that approximately one-third of the research literature on SJTs does not indicate the intended constructs measured or authors do not provide enough information about the constructs to reliably evaluate how well the relationships to other measures support the validity of SJT score interpretations. Of note, the research was limited due to feasibility constraints and the research questions to be addressed to investigating the relationship of SJT scores to a select number of variables.

Consequences of testing. *Standard* 1.25 (AERA, APA, & NCME, 2014) describes the need for evidence to address intended and unintended consequences of testing, which is particularly significant as SJTs are being used in the health professions to inform high-stakes decisions such as admissions or residency placement (Patterson et al., 2016). Cizek (2015) argues that consequences relate to validity evidence supporting the justification of test use, which is separate from validity evidence supporting interpretation of test results (i.e. the focus of this research). In his framework, he suggests validity evidence justifying test use can be derived from four sources, including consequences, alternative options, costs, and fairness.

In general, evidence of the consequences of using SJTs is limited. The most applicable research regarding SJTs as it pertains to consequences of testing is the impact of using SJTs to inform high-stakes decisions. For example, researchers have explored the extent to which students seek coaching to improve test taking strategies in addition to evaluations of how well

the responses can be faked (Lievens, Buyse, Sackett, & Connelly, 2012; Lievens, Peeters, & Schollaert, 2008; McDaniel & Nguyen, 2001; Whetzel & McDaniel, 2009;). Nguyen, Biderman, and McDaniel (2005) showed that SJTs tend to be more difficult for individuals who fake positive responses that would be more socially desirable depending on the item format. This quality makes SJTs highly favorable for use in admissions and selection decisions. Overall, the impact of testing on emotional well-being, influence on decision-making processes, and other consequences has not been developed in the literature but should be considered in future explorations.

Of note, the focus of the research was not to directly contribute to validity evidence related to consequences of test use defined by the *Standards* (AERA, APA, & NCME, 2014) or validity evidence for test use in general as desired by Cizek (2015). Evidence regarding the responses processes, however, may inform research agendas on the consequences of SJT testing if participants comment on potential consequences of performing poorly or their desire to provide positive responses.

Cognitive / response processes. *Standard* 1.12 (AERA, APA, & NCME, 2014) describes evidentiary sources for tests intended to measure cognitive or psychological processes. Of all the sources of validity evidence, the greatest void appears to exist regarding SJT response processes. For years, an understanding of SJT response processes has been a neglected area of research despite numerous requests from SJT researchers to contribute to the literature (Fan, Stuhlman, Chen, & Weng, 2016; Harris, Siedor, Fan, Listyg, & Carter, 2016; Melchers & Kleinmann, 2016; Ployhart, 2006; Sorrel et al., 2016). Understanding of SJT response processes is critical because SJTs are assumed to engage cognitive processes related to decision-making abilities and prioritization of actions, which has not been demonstrated empirically. Knowledge

of these processes can inform SJT design by identifying how design elements impact the response as well as awareness of individual attributes that may introduce construct-irrelevant variance into the score. In summary, there is little known about the response processes governing how individuals interact and respond to SJT items (Sorrel et al., 2016). This has been a neglected area of SJT for decades and serves as the primary focus for this research.

A review of the literature identified two studies that have reported on SJT response processes; however, both studies investigated elements of the response process as a minor component of their overarching research. As a result, these studies provide limited response process evidence for SJTs in general. First, a study by Krumm and colleagues (2015) aimed to identify the types of general domain knowledge test takers used when completing SJT items about teamwork without situational descriptors. Forty participants, including students and employees, were requested to think-aloud as they completed 18 SJT items that were designed to measure teamwork skills (e.g. conflict resolution, collaborative problem-solving, communication). The think-aloud interviews were coded to identify the strategies participants used to evaluate response options. It was hypothesized that participants would compare response alternatives or make a general evaluation of the response behavior to determine the best response. Of all the elicited statements during the think-aloud interviews, participants most often compared the response options (44.4%) in addition to evaluating the effectiveness of response options (40.2%). Their findings suggest test takers used the response options as a source of information, especially when there were insufficient situational descriptors provided in the stem of the item. A limitation, however, was that the study did not have SJT items with situational descriptors to explore how strategies may vary based on whether a descriptor is present.

Another study by Rockstuhl and colleagues (2015) used think-aloud procedures with 12 international managers in a multi-national study to describe the response process of multicultural SJTs. Participants were asked to think aloud about how they would respond to four SJT items that were presented as brief video vignettes. They discovered approximately 82% of comments about that SJT related to one of three categories: (1) intentions (e.g. what someone in the scenario wanted to do), (2) emotions (e.g. strong feelings about the situation), or, (3) thoughts (e.g. describing plans, actions, or ideas). The results were important as they identified what participants often thought about during the response process. The limitation, however, is that these elements were not combined to identify how this process was consistent across examinees (i.e. a consistent model of responding to SJT items) or how this process was aligned with the construct being measured (i.e. if the utterances suggested the test was tapping into the desired knowledge, skills, and abilities).

These studies are the only examples found in the literature that involved think aloud protocols or cognitive interviews to explore SJT response processes. Overall, these efforts have not been sufficient. As described, the focus of each study was very specific and did not significantly contribute to the holistic understanding of the process by which participants formulate their response to SJT items. Krumm's study (2015) was the best attempt in describing these processes compared to Rockstuhl (2015) who simply summarized the content of the utterances made by participants. There were no explicit connections to the theoretical underpinnings of the constructs being assessed or SJT methodology. The goal of the research proposed here, therefore, is to explore SJT response processes to contribute to this vital area of validity evidence.

Evaluating Response Processes in Assessments

Describing SJT response processes is a formidable challenge due a combination of poorly specified constructs and to the poor understanding of the cognitive processes engaged during the examination (Ployhart, 2006). The knowledge, skills, and abilities measured by an SJT are inherently complex; they include the integration and coordination of various practices, core concepts, as well as major ideas of the domain to determine the best response to a task or challenge (Nichols & Huff, 2017). Moreover, Ercikan and Pellegrino (2017) argue a critical reason for evaluating examinee response processes is to ensure the tasks tap into the intended knowledge and skills instead of assuming it occurs. Understanding participant response processes is, therefore, essential to interpreting scores from instruments intended to measure these abilities. This section outlines how examinee responses processes can be evaluated and offers a review of pertinent frameworks that will be applicable in analyzing SJT response processes.

During an assessment, an examinee activates a *cognitive response process*; this includes the moment-to-moment steps required to think and make decisions (Pellegrino, Chudowsky, & Glaser, 2001). An understanding of these processes is based on contemporary cognitive theories of learning which focus on how knowledge is organized and the procedures used for reasoning and decision making (National Research Council, 1999). The cognitive response process, therefore, includes how information is accessed, represented, revised, acquired, and stored to address a question. The decision-making process includes the manipulation of information in a series of steps, which can be informed by existing knowledge, experience with previous techniques, or the application of analogies; this process is also triggered by contextual cues. In general, cognitive response processes associated with specific schema are considered to be

domain-specific and, therefore, change depending on the setting (Pellegrino, Chudowsky, & Glaser, 2001).

Problem-solving processes include either weak methods or strong methods depending on the necessity of context. *Weak methods*, described by Newell and Simon (1972) are applicable in domain-general problem-solving processes. These can include procedures such as creating analogies or trial and error. Weak methods are important because they are often engaged when solving novel problems regardless of the level of expertise of the problem solver. Conversely, *strong methods* are applicable only in domain-specific problem-solving processes. These procedures include specific algorithms that pertain to a particular domain such as mathematics, scientific reasoning, or reading comprehension (Leighton & Gierl, 2011).

When it comes to assessing complex thought processes, evidence must demonstrate that test takers use cognitive processes in a coordinated fashion that is consistent with the theoretical and empirical expectations (Nichols & Huff, 2017). Evaluating cognitive response processes is often elaborate and can vary based on the context or the tasks being assessed. Evidentiary sources investigating cognitive response processes often include think-aloud procedures and cognitive interviews, each of which is used as part of an overall cognitive task analysis, in these cases to create verbal reports that can be annotated and analyzed to describe these response processes. Leighton (2017) outlines how each of these approaches can be used to explore as well as confirm cognitive response processes (see Figure 3).

A foundational perspective of assessing cognitive processes refers to research on cognitive aspects of survey methodology (Schwarz, 2007), which is also applicable to assessing SJT response processes because they both involve situating oneself in the context and choosing responses that would be guided by schema relevant in those situations. This approach considers

the task characteristics and respondent behaviors to describe the interplay between cognitive and communicative processes necessary for response to survey items. Tourangeau, Rips and Rasinski (2000) proposed a four-step process that participants use when completing a survey: (1) comprehension, (2) retrieval, (3) judgment, and (4) response selection. During *comprehension*, the examinee uses cognitive processes to read, interpret, and understand the purpose of the question. Next, the *retrieval* phase includes accessing long-term memories and knowledge relevant to the scenario and proposed problem. A *judgment* is formed by the examinee based on a complex integration of memories, knowledge, experiences, and other antecedents (Brooks & Highhouse, 2006). Finally, the examinee selects a *response* that is most consistent with their judgment.

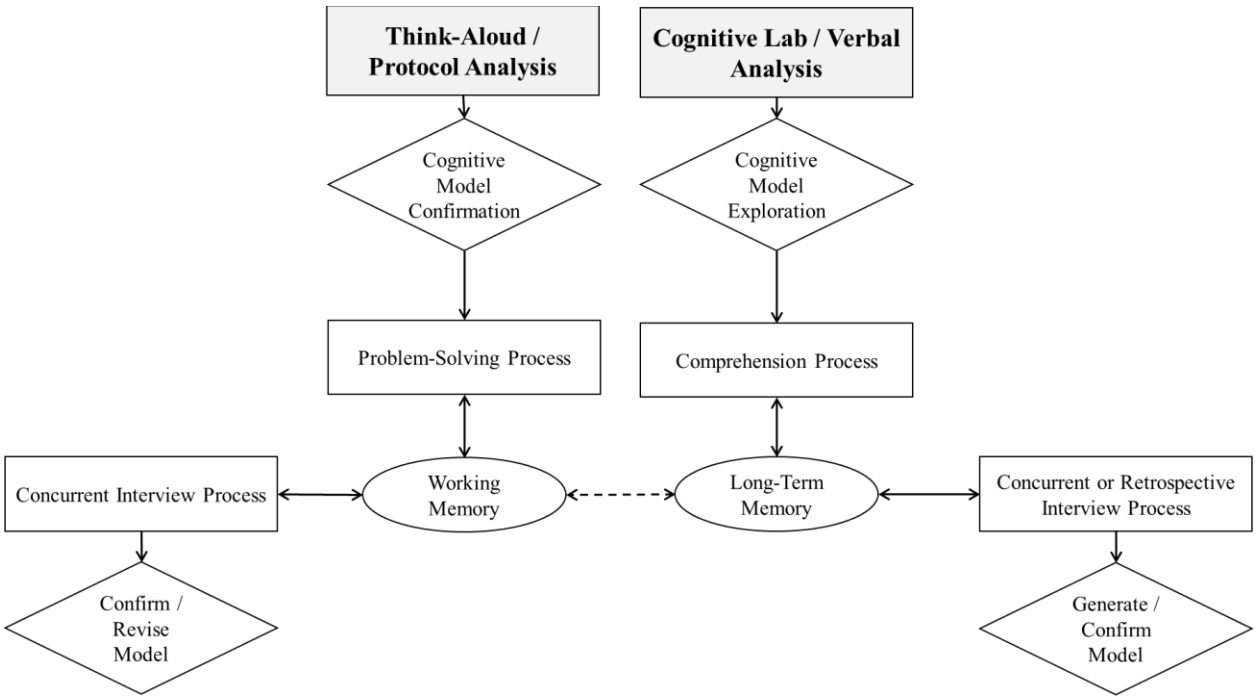


Figure 3. Differential measurement objectives for think-aloud interviews and cognitive laboratory interviews (adopted from Leighton, 2017)

Each item on a survey can be approached using this framework, which concludes with the participant making a judgment that is then mapped onto the pre-determined response options

for the best fit. Each stage can be affected by a variety of psychological mechanisms that can influence the final response. Ployhart (2006) proposed an SJT response model, shown in Figure 4, that added contextual factors specific to an SJT using the four-stages proposed by Tourangeau, Rips, and Rasinski (2000) as the foundation of this process. In general, he noted that sources of construct-irrelevant variance (such as language barriers, interpretation issues, and impression management) can affect all stages in addition to overall test-taking motivation (Ployhart, 2006). He argued, however, that certain elements were more likely to influence certain stages. He proposed, for example, participant knowledge contributes to each stage of the response process, but personality only influences the response selection in questions that are focused on what an examinee should do. In other words, personality may affect the entire process if the question asks what the examinee would do. Ployhart (2006) proposed that reading ability is significant for only the comprehension and response selection stages of written SJTs.

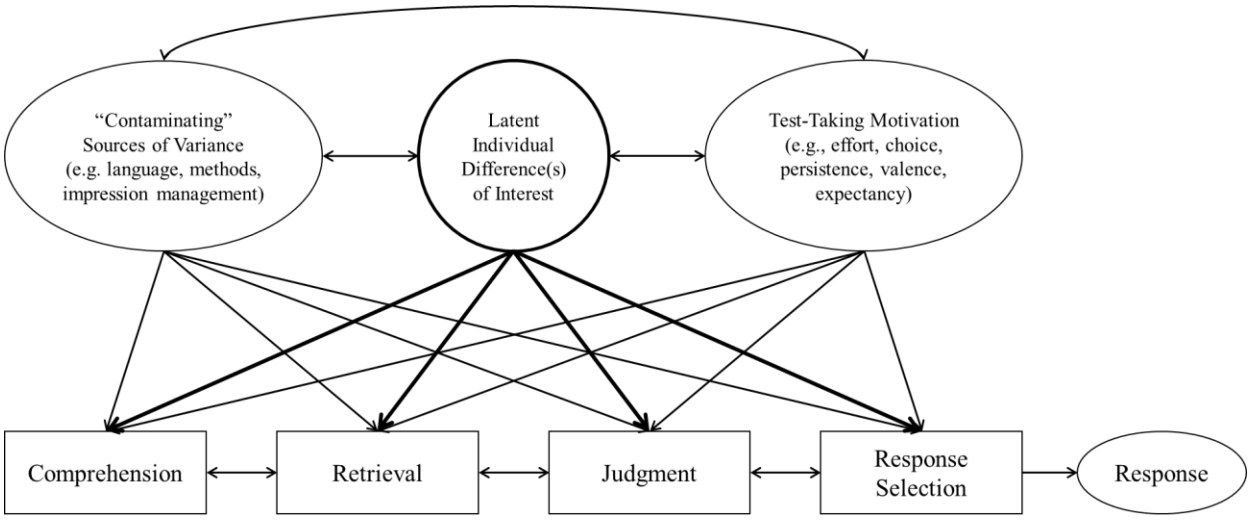


Figure 4. A model of SJT response processes (adopted from Ployhart, 2006)

The response process model proposed by Ployhart has not been fully tested. The relationship of the variables and attributes comprising the model are based on the relationships observed in research as well as hypothetical assumptions. Personality, for example, has been

shown to relate to SJT performance and this model attempts to identify where in the process it is suspected to have in influence (Ployhart, 2006). The purpose of the research proposed here is to evaluate whether these elements are salient in the cognitive processes engaged when completing an SJT as hypothesized by Ployhart. The four-stage model will serve as a framework that will be used when analyzing verbal reports; participants will be asked to reflect on the decision-making process and utterances will be coded according to the presence of the proposed four stages and response process.

Defining the Construct of Interest – Empathy

Describing SJT response processes cannot be entirely separated from the assessment of the construct of interest. In other words, it is anticipated utterances regarding the response processes will be highly connected to attributes and understanding of the construct; therefore, it is essential to provide a brief overview of the target construct: *empathy*. In addition, the review is intended to define the domain to be tested and will inform the construct-driven design of an SJT to ensure it aligns with our theoretical and empirical understanding of empathy.

Empathy in healthcare is an "elusive concept" (Hojat & Gonnella, 2015, p. 344). There is limited consensus about the definition of empathy or the salient factors despite decades of research across the health professions. Rogers (1951), a pioneer of client-centered counseling therapy, is often credited with the initial conceptualization of empathy in medical practice. Empathy, as he described it, was the "as if" (Rogers, 1951, p. 129); in other words, it was empathy that allows clinicians to understand a person's point of view, their feelings, and the potential causes of these perspectives and feelings. Hojat's (2007) definition of empathy is now commonly cited and will serve as the basis for this review and research. According to Hojat, "empathy is a predominantly *cognitive* (rather than an emotional) attribute that involves an

understanding (rather than feeling) of experiences, concerns, and perspectives of the patient, combined with a capacity to *communicate* this understanding and an *intention to help*" (p. 80).

Empathy is consistently considered to be a multidimensional construct that includes at least two factors: cognitive empathy and affective empathy (Hojat, 2007; Quince, Thiemann, Benson, & Hyde, 2016; Tamayo, Rizkalla, & Henderson, 2015). *Cognitive empathy* refers to an individual's ability to understand another person's perspective versus being self-oriented (Fjortoft, Van Winkle, & Hojat, 2011). This cognitive perspective includes being able to imagine alternative realities, to judge the difficulty of scenarios, and to "step into another person's shoes and to step back as easily into one's own shoes again when needed" (Hojat, 2007, p.8).

The other element, *affective empathy,* pertains to an individual's ability to understand and internalize the feelings experienced by others (Nunes, Williams, Sa, & Stevenson, 2011). Also called emotional empathy, affective empathy relates to recognizing the emotional response that can be generated by individuals or through the interactions between people (Hojat, 2007).

A third commonly accepted factor involved in empathy in the healthcare literature is behavioral empathy. *Behavioral empathy* consists of action-oriented responses that outwardly express the internally experienced cognitive and affective processes (Larson & Yao, 2005). Hojat's definition of empathy (2007) refers to behavioral empathy as the ability to communicate this understanding with others. The act of communicating explicates these thoughts and feelings, which can be instrumental for optimal patient care. Tamayo, Rizkalla, and Henderson (2015) believe the trinity of cognitive, affective, and behavioral empathy are necessary to practice patient-centered care; in other words, if any factor is lacking then care is not as effective.

A fourth factor, referred to as *moral empathy*, has been reported inconsistently in the literature. Moral empathy, defined by Morse and colleagues (1992), includes the internal

altruistic motivation to be empathic towards others. Subsequent studies have concluded that this factor is no longer a relevant feature of empathy (Decety & Jackson, 2004). Of note, Hojat's definition (2007) identifies moral empathy as the intent to help others.

For the purposes of this research, empathy was defined as having a two-factor structure with cognitive and affective elements. Although the three-factor structure (e.g., cognitive, affective, and behavioral) has been the basis of study in the health professions, empirical evidence of this factor structure is debatable (Quince, Thiemann, Benson, & Hyde, 2016). The *Jefferson Scale of Physician Empathy (JSE)*, for example, was designed to as a brief measure of health professionals' empathy (Hojat et al., 2001). The initial instrument included 90-items based on a thorough literature review and previously published instruments, such as the Interpersonal Reactivity Index (IRI). The 90-item instrument was reviewed by 55 physicians who provided feedback on the appropriateness of each item and wording based on the definition of empathy provided by the researchers. The revised instrument included 45 statements that participants evaluated on a scale of 1 (strongly disagree) to 7 (strongly agree). The revised instrument was completed by 41 internal medicine resident physicians and 193 third-year medical students at Thomas Jefferson University Hospital and Jefferson Medical College, respectively. A principal component analysis (PCA) with a varimax rotation was used to determine which items would be included in the final instrument as indicated by a factor structure coefficient greater than .40.

The results of the factor analysis identified a four-factor structure with one grand factor indicated by an eigenvalue of 10.64; of note the second factor had an eigenvalue of 3.45 and the other eigen values were not included. Twenty items were included in the final instrument with factor loadings ranging from .39 to .82. The four-factors included: (1) understanding the patient's perspective, (2) understanding the patient's experiences, feelings, and clues, (3) ignoring

emotion in patient care, and (4) thinking like the patient. Of note, 3 items cross-load onto multiple factors and some factors only included 2 or 3 items. Scores on the JSE also had weak, positive correlations with performance on the IRI, ranging from .24 to .40 (Hojat et al., 2001).

Additional psychometric analyses and validity evidence supporting the use of the JSE, however, has been limited. A second study by Hojat and LaNoue (2014) included response data from 2,637 medical students who completed the JSE at the beginning of medical school from 2002 to 2012. An exploratory factor analysis (EFA) was used to determine the factor structure for students from the 2002 to 2007 matriculating classes (n=1,380); this structure was then used to conduct a confirmatory factor analysis (CFA) on data from the 2008 to 2012 matriculating classes (n=1232). Three factors were identified labeled *perspective taking* (10 items), *compassionate care* (8 items), and a third undescribed factor (2 items) with eigenvalues of 4.7, 1.6, and 1.4, respectively. Factor loadings ranged from .29 to .75 with 2 items cross-loading on multiple factors. The CFA exemplified the 3-factor model had satisfactory fit (χ^2 (168, n = 1,232) = 887.87, $p < .001$, RMSEA = 0.05, TLI = 0.89), which does not conclusively support a 3-factor latent structure as anticipated. This study also included only matriculating students as opposed to also including those with more substantial amounts of practice experience.

In summary, although the JSE has been the standard approach to measure empathy in the health professions, a two-factor structure is more aligned with the current understanding of empathy according to evidence in the neurosciences, as described next; therefore, other instruments may be more appropriate to measure empathy and warrant further exploration (Carre, Stefaniak, D'Ambrosio, Bensalah, Besche-Richard, 2013; Gerdes, Segal, & Lietz, 2010).

The *Questionnaire of Cognitive and Affective Empathy* (QCAE), for example, is one instrument that was developed using items from 4 existing empathy instruments in the literature

and insights from the neuroscience (Reniers, Corcoran, Drake, Shryane, & Vollm, 2011). The initial QCAE included 65-items with 29 items related to cognitive empathy and 36 items related to affective empathy. For each statement, participants are asked their level of agreement on a scale of 1 (strongly disagree) to 4 (strongly agree) with higher scores related to higher levels of empathy.

Students and employees from the University of Manchester and Manchester Metropolitan University (n=640) completed the 65-item version of the QCAE, which was analyzed using a PCA. The PCA identified 10 factors with eigenvalues that exceeded 1; however, a scree test suggest only 5 factors were salient. Factor loadings suggested 31-items were appropriate to include on the final instrument with values ranging from .436 to .736. A CFA was conducted with a second sample of participants (n=318) to verify the 5-factor structure. The CFA exemplified the 5-factor model had satisfactory fit (χ^2 (80, n = 318) = 193.897, $p < .001$, RMSEA = .067 [90% CI (.055-.079)], CFI = .947, TLI = .930). In addition, scores on the QCAE were strongly correlated ($r = .62$, $p < .001$ for cognitive empathy; $r = .76$, $p < .001$ for affective empathy) with participant scores on the *Basic Empathy Scale* (BES), another recently developed instrument that measures cognitive and affective empathy (Jolliffe & Farrington, 2006). Of note, the BES was not included as a potential instrument as the QCAE includes a more comprehensive definition and assessment of components of empathy (Reniers et al., 2011).

Two of the factors of the QCAE are related to cognitive empathy (i.e., the ability to construct a working model of the emotional states of others) and three of the factors are related to affective empathy (i.e., the ability to be sensitive to and vicariously experience the feelings of others) according to the definitions created by Reniers and colleagues (2011). The sub-components of cognitive empathy include:

(1) *perspective taking* (10 items), which is defined as intuitively putting oneself in another person's shoes in order to see things from his or her perspective; and

(2) *online simulation* (9 items), which is the effortful attempt to put oneself in another person's position by imagining what that person is feeling and is likely to be used to consider the other person's future intentions.

The sub-components of affective empathy include:

(1) *emotion contagion* (4 items), which is defined as the automatic mirroring of the feelings of others;

(2) *proximal responsivity* (4 items), which includes the affective response when witnessing the mood of others in a close social context; and

(3) *peripheral responsibility* (4 items), which is the affective response when witnessing the mood of others in a detached social context such as a book or movie.

In summary, the two-component structure of empathy will be instrumental as a framework for generating SJT items and response options that are consistent with the theoretical definition of empathy. Items, for example, will be designed to address one of these two components (i.e., affective, cognitive) to obtain a holistic measurement of empathy that is theoretically based. Moreover, instruments like the QCAE can be used as a starting point to generate sample questions as each item in the survey is mapped to a specific empathy component. The empathy components will be a framework used when analyzing verbal reports. Utterances related to the construct of interest will be coded in reference to which component is being discussed.

In conclusion, empathy is an opportune construct to incorporate into this research due to its multifaceted nature and because it presents a realistic challenge for designing and evaluating

SJTs. Many of the constructs used to describe components of professional competence have ill-defined structures that make the process difficult. In addition, according to Quince and colleagues (2016), empathy is becoming as important as clinical competence in healthcare. This research, therefore, has practical implications as it offers a new strategy to evaluate empathy in the health professions.

Summary

The validity evidence to support the interpretation and use of SJT scores is mixed and generally inconclusive due to variability in SJT design and evaluation processes. Relationships to other constructs and criteria are the most studied; the available research has shown weak to moderate positive relationships with personality traits and cognitive ability. Overall, there is a need to study the response processes with SJTs to better understand the theoretical underpinnings of SJTs and contribute to a substantial void in the validity evidence for their use. A background in complex cognitive response processes can serve as a guide for analyzing SJT response processes. This research will focus on developing an SJT intended to measure the two components of empathy (e.g. cognitive and affective empathy), which are critical components of professional competence in the health professions.

Chapter 3: Methods

The purpose of this research study was to develop a greater understanding of the response processes involved in completing SJTs used in the health professions. This chapter includes a description of the instrument design, participants, and data collection, preparation, and analysis procedures.

Instrument Design

To evaluate the response processes during SJTs, an SJT was created to target a construct judged necessary for success in the health professions: empathy. In general, instrument development requires a comprehensive approach to ensure results contribute to assessing the construct of interest while minimizing construct-irrelevant variance. The process is frequently iterative and characterized by 12 critical components (Lane, Raymond, Haladyna, & Downing, 2016). Moreover, alignment with the *Standards* (AERA, APA, & NCME, 2014) provides a framework of evidence-based strategies consistent with best practices in the testing community. This approach from Lane and colleagues (2016) informed the design of the instrument so that this SJT would best approximate evidence-based design strategies used in practice; however, it is noted that not all steps were necessary or required to meet the exploratory research purposes of this project.

The first step—the overall plan—delineates the major activities involved in the development process and the validity evidence intended to support the score interpretations and

uses (Lane, Raymond, Haladyna, & Downing, 2016). The decisions made at this stage are based on current findings in the literature but are subject to change based on implementation.

Interpretation and use argument. The *Standards* (AERA, APA, & NCME, 2014) suggest sufficient evidence and theory must be provided to support the intended interpretations and uses of test scores; this is often completed using an argument-based approach to validation (Kane, 1992). This research focused exclusively on generating validity evidence supporting the intended interpretation of SJT scores; discussion about evidence supporting the use of an SJT for admission decisions or other purposes in this context is limited.

An essential goal of using an SJT as an instrument is to generate a score that is indicative of an examinee's standing on a targeted construct. For the purposes of this research, each item was designed to target one of the two subcomponents of empathy, the construct of interest. The overall score on an SJT was, therefore, representative of the unidimensional construct of empathy. The design of SJT items was based on Lievens' (2017) recommendation to use a construct-driven approach, which incorporates theoretical and empirical evidence to inform sound instrument design. The intended inference was that high scores on an SJT (i.e. examinee answers are most consistent with the keyed answers) were indicative of higher standing on the construct of interest (i.e. exhibiting more empathy), whereas low scores on an SJT were indicative of a lower standing on the construct of interest (i.e. exhibiting a lower degree of empathy).

Construct definition. According to the *Standards* (AERA, APA, & NCME, 2014), the construct to be tested must be "defined clearly and justified in terms of importance" (p. 181). The focus of most SJTs in the health professions has been on the broader concept of professional competency, which can be subdivided into a host of smaller constructs of interest as outlined by

Patterson and colleagues (2013). This SJT focused specifically on empathy as the pertinent construct of interest due to its significance in healthcare. Healthcare providers who are more empathic have been shown to contribute to positive patient outcomes (Kim, Kaplowitz, & Johnson, 2004; Reiss et al., 2008). As presented in chapter 2, empathy was defined for purposes of this study as the ability to understand a person's point of view and their feelings (Hojat, 2007).

Content specification. The *Standards* (AERA, APA, & NCME, 2014, Standard 11.1) recommend test content specifications identify the scope of the construct to be assessed and to describe test design features. With respect to SJT methodology, content specifications are critical as they define the framework for the scenarios that reflect job and practice experiences. To establish the content specifications for an SJT, a thorough analysis of the job and practice experiences is necessary and often uses multiple sources: organizational standards, theoretical frameworks and empirical evidence, and job/practice analyses (Patterson, Zibarras, & Ashworth, 2016).

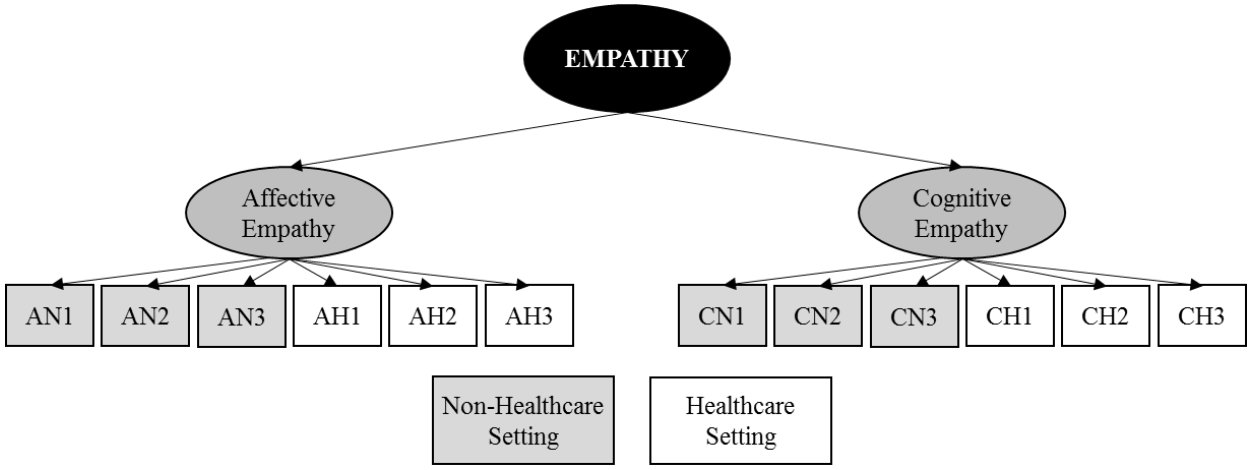


Figure 5. Map of SJT items, settings, and the associated construct components.

The SJT for this study was designed to target the two components of empathy as defined in chapter 2: (1) cognitive empathy and (2) affective empathy. The SJT for this study included 12 items with an equal number of items addressing only one of the two components of empathy (i.e.

6 items per component). Of the 6 items related to each component, 3 items were designed to address general domain knowledge (i.e. a non-healthcare setting), whereas the remaining 3 items were designed to incorporate job-specific knowledge (i.e. a healthcare setting). Figure 5 provides a visual representation of the item distribution and the assigned item label. The process for creating and selecting SJT items for this study is described later in the chapter.

Format specifications. SJTs are a unique assessment methodology as they can integrate various design formats depending on the targeted aims and objectives. A prominent focus in the literature has been identifying the design features that optimize validity and reliability data (McDaniel, Morgeson, Finnegan, Campion, & Braverman, 2001). This section addresses pertinent SJT design features used for this study and offers the rationale for their selection compared to alternatives.

Response instructions. For SJTs, the response instructions can influence the attributes being assessed with subsequent consequences for the validity and reliability of the results (McDaniel, Hartman, Whetzel, & Grubb, 2007; McDaniel & Nguyen, 2001). Questions can be tailored in either a behavioral tendency format (i.e. "how *would* you respond") or a knowledge format (i.e. "how *should* you respond").

The key difference between the two formats is that knowledge-based instructions (i.e. *should* do) are believed to be require job-specific knowledge and cognitive ability to select an accurate response. This is corroborated by evidence showing knowledge format questions are more correlated to cognitive ability tests and are less susceptible to faking than behavioral tendency formats (Nguyen, Biderman, & McDaniel, 2005). The diminished potential for faking relates to the necessity for job-specific knowledge or experiences that examinees cannot easily fake.

In addition, questions using the behavioral tendency (i.e. *would* do) format more often measure general knowledge that may not be specific to knowledge or skill sets required in a certain profession. Conversely, SJT questions with a knowledge-based format (i.e. *should* do) can reflect the maximal performance potential of an examinee; when asking what an examinee should do in a scenario, it does not limit their response to what the individual feels they would simply be able to do (Lievens, Peeters, & Schollaert, 2008). Based on these findings, SJT items developed for this study were structured using a knowledge-based format (i.e. *should* do) to ensure measurements of the constructs of interest include job-specific and general domain knowledge needed to succeed in pharmacy practice.

Response format. Unlike items assessing clinical knowledge, SJT items often have no definitive correct answer; instead, there are several responses to scenarios in practice that could be considered appropriate. This makes single-response item formats for SJTs less desirable for testing and requires a variety of other response formats to assess the target construct in a valid and reliable manner. Unfortunately, evidence regarding optimal design strategies and preferred response formats is lacking.

There are five main response strategies employed when designing SJTs, each with advantages and disadvantages (Weekley, Ployhart, & Holtz, 2006). Response instructions can request test takers to: (1) select the single-best response, (2) select the best and worst responses, (3) select multiple appropriate responses (usually 2-3 selections), (4) rank the desirability of responses relative to one another, and (5) rank the effectiveness of each option on a scale. Table 3 includes samples of select response formats. The consensus is that the format should be selected based on the scenario setting, the level of discrimination needed between examinees, the

necessity to identify certain response patterns, and the desired complexity based on the target population (Patterson, Zibarras, & Ashworth, 2016).

Table 3

Sample SJT Formats

Multiple Choice Format
1. A physician has asked for you to provide medication education regarding a new antidiabetic agent for a patient. You talk briefly with the patient, discuss important information about the medication, and prepare to leave. The patient appears worried. *Select the TWO most appropriate responses:* A. Allow the patient to share their concerns when they choose without directly asking. B. Tell the provider the patient appears to be concerned about the new medication. C. Speak with the patient to establish the possible concerns. D. Provide a handout with more detailed information about the medication for reference. E. Ask the nurse to ask if the patient has any particular concerns.
Ranking Response Format
1. One of your patients appears to be very depressed, which she believes to have been precipitated by the recent loss of a loved one. You realize her loss parallels one of your own experiences and wonder how this might be used to develop rapport with your patient. *Rank the following responses in order of 1=MOST appropriate to 5=LEAST appropriate.* A. Describe your own loss and subsequent feelings in detail. B. Acknowledge her understandable sadness from experiencing a personal loss. C. Change the subject, as dwelling on it may make her more upset. D. Encourage her to discuss her feelings with a friend, family member, or religious leader. E. Recommend she speak more with her provider about counseling services.
Rate Effectiveness of All Options
1. A patient at your hospital complains to you about how awful the hospital food has been. He mentions he saw a hot dog vendor during his admission to the hospital the other day and he has been craving one ever since. *For each response, rate the effectiveness of the response from 1 = NOT effective to 5 = VERY effective* A. Ask the patient what in particular he has disliked about the food. B. Agree with the patient that the hospital food is not the best. C. Get a hot dog from the vendor across the street for the patient. D. Tell the physician the patient would like better food options if they are available. E. Contact the food services team and file a complaint about the food on behalf of the patient.

Ranking response options, for example, is ideal when prioritization of tasks is to be measured and when faking is to be minimized due to the complexity of the task. Ranking response options also allows for partial credit compared to other response formats; therefore, it is preferred if greater granularity of individual performance is desired, such as in high-stakes selection. A disadvantage of ranking response options, however, is the increase in cognitive load and time necessary to complete each question; therefore, time constraints must be considered.

Response formats that request the test taker to select multiple responses are more useful in scenarios where the order of activities is not essential, but completeness is. For example, if a response to a scenario requires multiple actions in a non-critical sequence, a format that allows the test taker to select the most applicable options is appropriate. In cases where knowledge of what *not* to do is essential, examples with the best and worst selections would be warranted to ensure the distinction in their knowledge is clear. Single-response options in SJTs are emerging as potential options; however, the applicability to the health professions has not been extensively evaluated (Crook et al., 2011; Motowidlo, Crook, Kell, & Naemi, 2009).

With respect to the impact on reliability, St-Sauveur and colleagues (2014) found the single-response option to provide the lowest internal consistency compared to rank ordering and best/worst response formats. Ployhart and Ehrhart (2003) showed that rating the effectiveness of each response option results in the highest internal consistency and the single-response option was the lowest. Relationships between response formats and validity could not be identified in the literature search.

In current high-stakes testing programs, a combination of response formats is recommended to balance feasibility and desired outcomes (Goss et al., 2017; Patterson, Zibarras, & Ashworth, 2016;). The Foundation Programme SJT, for example, has half of the items as

multiple-response format (e.g. select two of the most appropriate options out of five) and the remaining half as ranking-response format (e.g. rank the five options in order of appropriateness). The mixture allows for ranking questions to be used when prioritization may be necessary and to obtain finer granularity in individual performance while also being cognizant of testing time as multiple-response formats will not take as long to complete. Although it was preferred to include varying response formats in this study, only one format was selected for practical and logistic reasons. Therefore, all SJT items used in this research were ranking-response formats as this is the broadest response format and requires participants to analyze and discriminate among all options for each item. The ranking-response format requires participants to be more explicit in their decision-making processes, which offered a distinct advantage for this research compared to other response formats.

Test length and time. According to *Standard* 4.14 (AERA, APA, & NCME, 2014), test length and time must be evaluated and determine if a speed component is appropriate. For the purposes of this SJT, speed is not a necessary component of the construct of interest. Based on examples in the literature, approximately 2 minutes per question is sufficient for establishing time constraints (Christian, Edwards, & Bradley, 2010; Goss et al., 2017; Patterson, Zibarras, & Ashworth, 2016). In this research, participants were allowed as much time as desired.

Item development. Consistent with evidence-centered design principles, SJT items for this research project were carefully constructed and systematically selected to minimize construct-irrelevant variance as recommended by test development experts (Lane, Raymond, Haladyna, & Downing, 2016). For example, test development followed the evidence-centered design process to clearly define the construct of interest, recruit subject matter experts with a diverse range of experiences, and include a systematic process to evaluate the appropriateness of

each item at measuring the targeted construct based on subject matter expert opinion. The following section outlines the strategies and procedures used to develop items to ensure they targeted the construct of interest to the greatest extent.

Subject matter expert recruitment. All SJT items were developed and reviewed by subject matter experts, which consisted of pharmacy faculty and practitioners. A sampling frame of 30 individuals was constructed to include those who are frequently involved in assessment initiatives at the UNC Eshelman School of Pharmacy (e.g., members of assessment committees, faculty who assist with admissions interviews, and practitioners who teach students in the classroom and in practice settings). In summary, the frame included a convenience sample of faculty and practitioners that would be readily accessible and able to attend item development and review sessions. In addition, the list was compiled to include individuals from multiple practice settings (e.g., academia, research, ambulatory care, community, and hospital settings).

The 30 pharmacists were contacted directly via email and requested to participate in the SJT development process with an outline of the expectations. The goal was to recruit a total of ten individuals to serve as either item writers or reviewers; 11 individuals agreed to participate and attended one of two workshops to either write or review the items based on their availability. Subject matter experts also completed the QCAE and a brief demographic survey—the results are provided in Tables 4a and 4b. In addition, the subject matter experts were asked to provide feedback to optimize the demographic survey, which was also used to collect information from study participants. As seen in Tables 4a and 4b, the subject matter expert groups were, overall, highly trained and experienced in pharmacy practice in various clinical settings.

Table 4a

Subject Matter Expert Demographics (Group 1 N=7; Group 2 N=4))

	Group 1: Item Design n (%)	Group 2: Item Review n (%)
Female	3 (43)	3 (75)
Education and training		
Doctor of Pharmacy	7 (100)	4 (100)
Residency (e.g. PGY1 and/or PGY2)	3 (43)	2 (50)
Fellowship or post-doc	6 (88)	2 (50)
Advanced degree (e.g. MPH, MBA, PhD)	4 (57)	0 (0)
Board certification (e.g. BCPS, BCOP)	1 (14)	1 (25)
Practice area		
Ambulatory care	2 (29)	3 (75)
Cardiology	0 (0)	1 (25)
Global health	1 (14)	0 (0)
Oncology	1 (14)	0 (0)
Pediatrics	1 (14)	0 (0)
Research	2 (29)	0 (0)
Self-reported training related to empathy	6 (86)	1 (25)

Table 4b

Subject Matter Expert Demographics (Group 1 N=7; Group 2 N=4)

	Group 1: Item Design Mean (Range)	Group 2: Item Review Mean (Range)
Years licensed as a pharmacist	9 (4–26)	18 (5–26)
Years with a health professions faculty appointment	5 (1–23)	7 (2–18)
Average number of hours working in a healthcare setting per week	21 (0–65)	13 (0–45)
Average number of patients interacting with per week	7 (0–24)	1 (0–4)
Average number of students interacting with per week	18 (1–45)	3 (0–10)
Average number of non-pharmacist healthcare providers interacting with per week	5 (1–10)	2 (0–4)
Years of work experience in non-healthcare-related human services field	5 (0–18)	7 (4–10)
Questionnaire of Cognitive and Affective Empathy (QCAE) Score	88 (77–99)	90 (81–94)
Cognitive Empathy (CE) Score	57 (52–62)	61 (52–68)
Affective Empathy (AE) Score	31 (19–44)	29 (26–32)

Item writing. The researcher coordinated a small item writing workshop with the first group of seven subject matter experts. The purpose of the workshop was to create twenty-four SJT items to be reviewed and refined by the second group of subject matter experts. During the session, participants were provided a handout (Appendix A) that outlined the research questions, defined empathy, and provided instructions for the session.

Subject matter experts were divided into four groups and assigned one of the two subcomponents (i.e. cognitive and affective empathy). Each group was instructed to create 6 test items that included five plausible response options each. Three of the six items were required to be in a healthcare setting and the remaining three were required to be in a non-healthcare setting. In addition, the groups were requested to submit a proposed key for each item they created.

With regard to item content, this SJT was intended to reflect scenarios that are plausible in pharmacy practice; therefore, attention was paid to the content and response options for each item. During the workshop, subject matter experts utilized information from published literature, practice analyses, personal experiences, sample SJT items, and theoretical constructs to guide the development of each item. Currently, there is no robust evidence to suggest one source for content is preferred or has benefits psychometrically. Evidence does suggest that test takers prefer cases that are relevant and applicable to the construct being assessed (Clevenger et al., 2001); therefore, it was highly recommended situations be based on real events.

It was requested that item content relate to the experiences of practicing clinicians and be inclusive of various practices settings, such as within the hospital and ambulatory care pharmacy sites. A common dilemma in health professions SJT development is the desire to include in-depth clinical knowledge pertaining to the scenario (Patterson, Zibarras, & Ashworth, 2016). It was stressed, however, that assessment of clinical knowledge is outside of the scope of SJTs and that the focus was to be exclusively on the construct of interest (i.e. empathy). Clinical information was minimized unless it included pertinent job-specific knowledge that was necessary to identify an appropriate response. In summary, items were designed to target attributes of empathy with a balance of items incorporating job-specific knowledge (e.g. taking

place within a health care setting) and general domain knowledge (e.g. interactions with people or scenarios outside of the health professions).

During the workshop, participants were also instructed on the best approach to create the item structure. In most cases, SJT questions are structured in a three-part framework of antecedent-behavior-consequence (Weekley, Ployhart, & Holtz, 2006). The antecedent describes what led up to the situation, followed by the behavior which describes what the person did, and then concludes with the consequence related to the person's behavior. In some items, the antecedent may be the only component of the stem and the behavior is located within the response. This structure was recommended as a starting point for the test writing process.

Item review. A second workshop was organized with four different subject matter experts; the goal of this session was to revise and evaluate each of the 24 draft questions that were created by the first group of subject matter experts. Prior to the workshop, the questions were reviewed by the researcher and edited for grammar and complexity. During the second session, participants were provided a handout (Appendix B) that outlined the research questions, defined empathy, and provided instructions for the session.

Subject matter experts in the second session were instructed to complete the pilot SJT independently, which included ranking each of the response options from most (1) to least (5) appropriate based on how they should respond to the provided scenario. In addition, they were asked to evaluate how well each item measured empathy on a scale of 1 (*Very Poorly*) to 5 (*Very Well*). Participants were also requested to identify if they believed the item addressed affective or cognitive empathy and to distinguish if it included a healthcare or non-healthcare setting. Lastly, they were requested to provide feedback and revisions regarding any SJT items. Participants also completed the QCAE and a demographic survey—these results of which are included in Tables

4a and 4b. A fifth subject matter expert also completed the review in the event of a tie when evaluating the items for selection. Data from this individual were not included. Overall, there were minimal changes to item wording or structure during the second session.

Item selection. The response and evaluation data from the second session were aggregated to determine which items would be included in the final SJT based on a set of pre-determined decision criteria described more explicitly in the next paragraph. In summary, SJTs items were included in the final test if there was a high level of agreement among subject matter experts on the ranking of response options (i.e. the rational key), if the item was perceived to be a good measure of empathy, and if there was majority agreement (e.g., at least 3 of the 4 reviewers agreed) that it measured the intended subcomponent of empathy (i.e. cognitive or affective empathy in a healthcare or non-healthcare setting).

To evaluate the level of agreement on the ranking of responses, the rankings from the four subject matter experts were compiled and Kendall's coefficient of concordance was calculated. A Kendall's coefficient value of .6 or above is a preferable level of rater agreement (Patterson et al., 2009; Siegel & Castellan, 1998); therefore, any items with a coefficient less than .6 were excluded. Three items were excluded due to poor concordance among the subject matter experts with values of .54, .55, and .57, respectively.

To evaluate whether each item was a good measure of empathy, the subject matter experts rated each on a scale from 1 to 5. An Empathy Index was computed as the average of these ratings; mean values on the Empathy Index less than 3.0 were considered to indicate that an item was a weak measure of empathy according to participants. Six items were excluded with mean ratings from 1.50 to 2.75.

To confirm that items were mapped appropriately to the respective subcomponents of empathy, subject matter experts judged each item as measuring the affective or cognitive component. Items were identified as appropriately mapped if three of the four subject matter experts agreed with the initial designation determined by the first group of participants. A Component Index capturing the agreement between subject matter expert judgments and initial item designations was computed as follows. For items that were initially designated as measuring the affective component of empathy, the item received a value of 1 if a rater judged the item to be measuring the affective component (i.e., a rater agreed with the initial designation); the item received a value of 0 if a rater judged it to be measuring the cognitive component (i.e., a rater disagreed with the initial designation).

Likewise, for items that had an initial designation as measuring the cognitive component of empathy, the item received a value of 1 if a rater judged the item to be measuring the cognitive component (i.e., agreement); the item received a value of 0 if a rater judged it to be measuring the affective component (i.e., disagreement). Mean values of the Component Index were calculated and are reported in Table 5; mean values closer to 1.0 indicate greater agreement with the initial component designation and values closer to 0.0 indicate greater disagreement with the initial component designation. Items were excluded if the Component Index was .5 or less; as a result, three additional questions were excluded from the initial pool due to disagreement about the subcomponent being assessed.

Finally, to evaluate whether the focus of an items was a healthcare setting or non-healthcare setting, a Setting Index was computed. The subject matter experts assigned a value to 1 to items that they judged as having a healthcare setting focus and 0 to items they judged to

have a non-healthcare setting focus. Mean Setting Index values were calculated across raters and are reported in Table 5. No items were excluded based on the Setting Index.

A summary of the item evaluation criteria is provided in Table 5. In addition, the intended focus of the question as determined by the first group is given for reference. Items removed from further consideration for the study because of failure to meet one or more of the criteria are indicated using bold type for the criterion that caused the item to be rejected. The final situational judgment test (Appendix C) included 12 items that were equally distributed according to subcomponent and setting. A summary of the test item content is provided in Table 6 for reference.

Table 5

SJT Item Evaluation Criteria Based on the Subject Matter Experts

Item Number	Intended Focus	Concordance Coefficient	Empathy Index	Component Index	Setting Index	Final Item Label
1	AH	0.84	**2.50**	**0.25**	1.00	***
2	AH	0.96	3.25	0.75	1.00	AH2
3	AN	0.86	3.25	**0.25**	0.00	***
4	AH	0.91	**2.00**	**0.25**	1.00	***
5	AN	0.86	4.25	0.75	0.00	AN3
6	AH	0.96	4.25	0.75	1.00	AH3
7	AN	0.97	3.50	**0.50**	0.00	***
8	AN	0.68	**2.50**	**0.50**	0.00	***
9	CH	0.92	3.25	**0.50**	1.00	***
10	AH	**0.54**	3.75	**0.25**	1.00	***
11	AN	0.74	4.50	0.75	0.00	AN2
12	AH	0.91	4.50	0.75	1.00	AH1
13	CH	0.90	3.25	0.75	1.00	CH1
14	AN	0.65	3.00	0.75	0.00	AN1
15	CN	0.94	**1.50**	**0.50**	0.00	***
16	CN	**0.55**	4.50	0.75	0.00	***
17	CH	0.76	3.25	0.75	1.00	CH2
18	CN	**0.57**	3.75	0.75	0.25	***
19	CN	0.62	3.25	0.75	0.00	CN2
20	CH	0.89	3.00	0.75	1.00	CH3
21	CN	0.71	3.75	0.75	0.00	CN3
22	CN	0.79	**2.75**	0.75	0.00	***
23	CN	0.96	4.00	0.75	0.00	CN1
24	CH	0.64	**2.00**	0.75	1.00	***

Notes: *** = Item omitted from final pool due to failure to meet one or more criteria.
Final Three Character Item Label Key -- First Character: A = Affective; C = Cognitive; Second Character: H = Healthcare Focus, N = Non-Healthcare Focus; Third Character: 1, 2, 3 = Item Number

Table 6

Summary of SJT Item Content

Item Label	Subcomponent	Setting	Item Summary
CH1	Cognitive	Healthcare	A patient complains that the doctor never listens to them
CH2	Cognitive	Healthcare	Trouble getting a medication history from a pharmacist
CH3	Cognitive	Healthcare	Suspect a patient is lying about their diabetes management
CN1	Cognitive	Non-healthcare	A friend is going to use medications to help them study
CN2	Cognitive	Non-healthcare	A woman asks you to cut in line at a store when you're late
CN3	Cognitive	Non-healthcare	Your family questions your sibling's relationship status
AH1	Affective	Healthcare	A patient discusses the recent loss of a loved one
AH2	Affective	Healthcare	A nurse asks you to discuss a medication error with family
AH3	Affective	Healthcare	A family gets upset while you review their chemotherapy
AN1	Affective	Non-healthcare	A parent quickly becomes upset at a grocery store
AN2	Affective	Non-healthcare	A relative is upset about difficulty conceiving
AN3	Affective	Non-healthcare	A best friend is visiting and planning to drop out of college

Participants and Recruitment

The aim of this research was to describe the response processes used by examinees completing SJTs in pharmacy practice. The researcher recruited participants from two levels of experience: (1) student pharmacists (i.e. individuals completing their Doctor of Pharmacy) and (2) experienced practitioners (i.e. those with more than 5 years of experience as a licensed pharmacist). The purpose of the two levels of experience was to explore how differences in job experiences may influence SJT response processes. For example, it is unknown how much examinees draw on their prior experiences in selecting a response option during an SJT. Evaluating the cognitive processes of clinicians with a different degree of experience has the potential to identify key differences that can inform SJT design, scoring methods, or how to develop the construct of interest in novice learners.

The sample size necessary to evaluate response processes of surveys, instruments, and tests using qualitative methods varies according to the study objectives (Leighton, 2017; Willis, 2015). The primary purpose of this research was exploratory, with minimal plans to compare quantitative measures from the reports. Therefore, a sample of 15 participants was used for each of the study groups. This number was based on previous work that suggests 11 participants is

sufficient to achieve saturation of coding schemes when investigating SJT response processes (Rockstuhl et al., 2015).

According to Keppel (1991), a sample size of 17 participants per group has 80% power to detect a large effect size (d = .8) with 5% Type I error rate whereas a sample of 44 participants per group has similar error rates to detect a moderate effect size (d = .5). To date, there is no published evidence to suggest what type of differences in SJT performance may be exhibited between individuals with varying degrees of experience; therefore, a sample of 15 individuals per group was considered adequate to detect differences.

Participants were recruited using convenience sampling; individuals were contacted through local networks and personal contacts requesting their participation. Student pharmacists from all actively enrolled classes (second, third, and fourth years) were recruited from the University of North Carolina (UNC) Eshelman School of Pharmacy through email (see Appendix D). The goal was to have as equal of a distribution as possible among the three classes (i.e. 5 students from each class). Students were offered incentives for participating, including the chance to win one of two $25 Amazon™ gift cards.

Experienced practitioners were recruited through the UNC Eshelman School of Pharmacy preceptor listserv and personal networks (see Appendix B). Moreover, pharmacists located in the Chapel Hill and Durham, North Carolina area were specifically targeted and requested to participate. Pharmacists were not provided any incentives for participating in the study. The goal was to recruit a diverse collection of pharmacists with varying clinical expertise and experiences in pharmacy practice.

Table 7

Summary of Data Collection and Data Analysis Procedures with Associated Research Questions

Data Collection Technique & Description	Research Question	Data Coding & Analysis
Think-Aloud Interview Participants think aloud with minimal prompting by the interviewer while they complete each item	RQ1 (cognitive process & strategies)	*Coding*: Transcripts will be analyzed using the codebook to identify the frequency of major codes. The goal is to identify what elements examinees most often describe as it relates the decision-making process without prompting. *Analysis*: Prevalence and patterns of themes consistent with the theoretical models of SJTs and a comparison of the distribution of the codes in cognitive interviews and think-aloud interviews.
Cognitive Interview Participants review each item and are asked how they decided to rank each response option	RQ1 (cognitive process & strategies)	*Coding*: Transcripts analyzed using the codebook to identify the frequency of major codes. The goal is to identify what elements examinees most often describe as it relates the decision-making process. *Analysis*: Prevalence and patterns of themes consistent with the theoretical models of SJTs and a comparison of the distribution of the codes in cognitive interviews and think-aloud interviews.
Cognitive Interview Participants review each item and are asked what experiences they may have thought about when answering the question	RQ2 (role of job-specific experience)	*Coding*: Transcripts analyzed using the codebook to identify the frequency of major codes. The goal is to identify if job-specific experiences are recalled in the response process and if the frequency of codes based on the type of question and examinee. *Analysis*: Identify and describe the types of experiences recalled during the SJT and potential differences between novice and experienced clinicians.
Cognitive Interview Participants review each item and are asked how the context was important in answering the question	RQ3 (role of item setting)	*Coding*: Transcripts analyzed using the codebook to identify the frequency of major codes. The goal is to identify if examinees are attentive to the context presented in the question and the frequency of codes based on the type of question and examinee. *Analysis*: Identify and describe the how the context was perceived to influence the decision-making process and if the distribution of codes differed between items of different contexts.
Cognitive Interview Participants review each item and are asked what they think the question was assessing	RQ4 (role of ability to identify construct)	*Coding*: Transcripts analyzed using the codebook to identify the frequency of major codes. The goal is to describe how often examinees indicated the construct being assessed related to empathy. *Analysis*: Identify and describe which constructs the examinees believe is being assessed and how that distribution relates to elements of the item.

Data Collection Procedures

Data collection was organized to address either examinee performance or to address SJT response processes. A summary of the data collection techniques and a brief description is provided in Table 7. Of note, the focus of this research was to describe SJT response processes; therefore, the research questions have been mapped accordingly onto these techniques. The additional data related to examinee performance was intended only to describe how well the instrument performed.

Participation and consent. Those who agreed to participate in the study were invited to complete a 90-minute one-on-one interview with the researcher. The time for the interview was established based on a small-scale pilot; two pharmacists were requested to complete the full SJT in addition to a mock cognitive interview to estimate the time it would take to complete each component. The 90-minute selection allowed for sufficient time to complete this SJT followed by extensive questioning about eight SJT items. This time would also be feasible for practicing pharmacists who were recruited to participate in the study while minimizing any disruption to their workflow.

The researcher hosted interviews at practice sites (e.g. hospitals, clinics, community pharmacies) whenever feasible to encourage participation in the study. At the beginning of the interview, all participants were notified of the risks associated with the study and they were required to provide voluntary consent consistent with the requirements of the Institutional Review Board (see Appendix E). Individuals had the right to discontinue their participation in the study at any point and any collected data from that interaction would be destroyed and excluded from further analysis. Consent into the study also included the authorization to audio record the interaction for analysis purposes.

Test and survey administration. The 12-item SJT (Appendix C) was administered to each participant on paper that was labeled with a randomly assigned participant identifier (P01-P15 for pharmacists and S01-S15 for students). The paper administration of this SJT allowed the researcher to readily organize the questions differently for each participant to minimize order effects of questions. Appendix F provides a summary of the order in which participants received each test question. The paper administration also allowed participants to easily review their responses during the cognitive interview conducted after they completed this SJT.

During the examination, participants were asked to complete one test item at a time; they were not allowed to revisit prior questions once they had submitted their answer, which is consistent with SJT formats in the literature (Patterson, Zibarras, & Ashworth, 2016). The researcher attempted to create a standardized testing environment for all participants, which include minimizing distractions. Participants were not allowed to start the test until explicitly instructed by the researcher. At the conclusion of the session, participants completed the QCAE (Appendix G) and a brief demographic survey that was specific to either students (Appendix H) or pharmacists (Appendix I). All surveys were labeled with their unique participant identifier.

Think-aloud interviews. Each recorded interview began with an overview of the think-aloud procedures and expectations for this session. Instructions were crafted to minimize potential sources of bias that could be introduced by the interviewer. During a think-aloud, for example, the interviewer should explicitly state the purposes of the research. In this case, that included assuring the examinee that the research was exploratory with the intent of informing the design of future tests. If possible, it is recommended that the interviewer not be an expert in the construct being studied and the interviewer should state so as an approach to minimize anxiety that can be induced in think-aloud interviews (Leighton, 2017). In this research, the interviewer

is considered an expert; therefore, the goal was to emphasize the exploratory nature of the research to reduce any potential anxiety as much as possible. A script was constructed that outlined the participant's purpose in the research and how participants should conduct a think-aloud (see Appendix J).

During the think-aloud, participants were instructed to verbalize their thoughts as they worked through SJT items; the interviewer was only to intervene in the event of silence lasting greater than five seconds and could only use prompts such as "keep talking" (Leighton, 2017). The addition of prompts such as "what are you thinking" has been shown to affect cognitive processes that elicit elaboration and comprehension, which detracts from the purpose of the think-aloud to capture the problem-solving process. The participants completed all 12 SJT items uninterrupted by the interviewer during the think-aloud process, unless they were silent for a prolonged period. Participants could take a short break following the completion of the think-aloud interview prior to beginning the cognitive laboratory interview.

Cognitive laboratory interviews. Following the think-aloud interview, participants began the cognitive laboratory interview, which focused on their understanding of and approach to SJT items. The cognitive interview is reserved for *after* the think-aloud interview as requesting individuals to elaborate and describe their approach has been shown to alter cognitive processes (Chi, 1997). Reserving this approach until after the think-aloud interview protects against introducing biases into participant thought processes. Similar to the think-aloud procedures, prior to starting the cognitive interview the researcher discussed the process of the interview and expectations for this process following an explicit script (see Appendix K).

The distinct difference between the think-aloud and cognitive interview is that the cognitive interview included questions related to how participants solved each problem and why

they made certain selection decisions. Participants had the opportunity to review each item and their responses as they answered the cognitive interview questions. However, participants were not permitted to change their submitted responses.

The interview protocol (Appendix K) included a series of questions intended to address the research questions previously described. In addition, the questions were pre-determined to ensure consistency across participant responses. Consistent with cognitive interviewing techniques, further probe questions and alternative phrasing were provided (Leighton, 2017; Willis, 2015). The researcher also had the opportunity to ask additional questions, if time permitted.

The aim of the cognitive laboratory interview was to gain insight into the role of attributes considered to be relevant in decision-making processes during SJTs. The theory posited by Lievens and Motowidlo (2016) suggests factors that influence SJT performance can include: values, interests, personality, and emotional intelligence as well as general and job-specific experiences. The cognitive interview was designed to include questions that probe whether these attributes had a significant contribution to their selection of responses. Other factors such as impression management and the ability to identify the construct being tested are suspected to influence SJT performance (Griffin, 2014); therefore, questions were included to target the influence of these attributes.

Due to time constraints, participants were not asked to evaluate their responses for all 12 SJT items. Instead, each participant was asked about their responses to eight items, which was feasible in the 90-minute interview schedule. Based on the participant identifier, individuals were assigned eight items that were evenly distributed based on the subcomponent of empathy assessed and the setting. In other words, participants reviewed four items in a healthcare setting,

four items in a non-healthcare setting, four items measuring cognitive empathy, and four items measuring affective empathy to varying degrees of overlap. In summary, the 12 SJT items were assigned so that there were twenty cognitive interviews conducted per item including ten interviews with students and ten interviews with pharmacists. A summary of this distribution is provided in Appendix F. The cognitive interviews concluded with the distribution of the QCAE and demographic survey.

Data Preparation Procedures

All data were compiled and stored on a secure drive that was accessible only by the researchers. Data collection included audio data from the interviews, notes created by the interviewer, and the response data collected on paper for this SJT, QCAE, and demographic survey. All participants were given a randomly assigned unique identifier for data analysis procedures to ensure anonymity. Records of the key linking the participant name to the identification number were destroyed following data preparation. All data preparation techniques were consistent with Institutional Review Board requirements.

SJT performance data and survey responses. Participant SJT responses (e.g. rankings), QCAE, and demographic survey responses were recorded on the paper provided to examinees. All data were converted and stored in an electronic database using Microsoft Excel™. Responses were labeled using the participant identifier and no other distinguishing information to protect participant anonymity.

Participant responses to this SJT were recorded so that each response option was assigned a numeric value (i.e. 1 through 5) corresponding to which response they thought was best (i.e. a value of 1) and worst (i.e. a value of 5). Responses to the QCAE and demographic surveys were also coded based on the numeric values provided and values assigned to distinct categories

created by the researcher. The data file was reviewed to check for missing data and the presence of errors, such as a duplicate or tied rankings. The data were stored on a secure drive that was only accessible by the researchers.

Interview data. Audio files from the interviews were converted to written transcripts using an Institutional Review Board approved online transcription service. No additional information was provided to the transcription service that could identify the participants to ensure participant confidentiality was maintained. All efforts were made to not refer to a participant by name during the audio recording to ensure anonymity. The researcher reviewed the final transcripts to confirm their accuracy, correct discrepancies, and remove potential participant identifiers.

The de-identified transcripts were segmented in various ways to optimize data analysis procedures. For the think-aloud interviews, the entire interview was maintained in its presented order and grouped by the level of the participant (i.e. student or pharmacist). For the cognitive interviews, the segments were grouped according to the test item. For example, all cognitive interview questions related to item CH1 were grouped into one transcript for analysis and subdivided based on whether it was a student or a pharmacist. The de-identified transcripts were stored on a shared drive that was only accessible by the researchers. The notes about observations created by the researcher during the interview were also accessible to other researchers assisting with the study. Of note, these artifacts were not intended to be used as critical elements of the research process but may inform future SJT studies.

Data Analysis Procedures

Data analysis consisted of four distinct phases: (1) an analysis of the demographic and QCAE data, (2) an analysis of SJT performance data, (3) an analysis of the cognitive laboratory

interviews, and (4) an analysis of the think-aloud interviews. The primary focus of the research was on SJT response processes; therefore, SJT performance data analysis was included to provide an understanding of the psychometric qualities of the instrument prior to an in-depth qualitative analysis. This initial step was also necessary to evaluate the quality of the data and provide insights that would help explain response processes described in the cognitive and think-aloud interviews.

The following section outlines the sequential data analysis procedures conducted during the study. One item of attention is the order in which the data were analyzed. Specifically, the cognitive laboratory interviews were analyzed *prior* to the think-aloud interviews, which is contrary to the order in which the data were collected (i.e. participants completed the think-aloud prior to the cognitive interview).

The think-aloud interview was conducted prior to the cognitive interview during the study to not bias the response process that occurred naturally while participants completed this SJT. For data analysis, however, the cognitive interviews were anticipated to provide richer details about the cognitive processes or antecedents that may not be explicitly stated in the think-aloud; therefore, qualitative analysis of the cognitive interviews prior to the think-aloud allowed the researcher to create a more robust for analyzing the think-aloud interviews. This is beneficial as it could identify if certain strategies or processes highlighted in the cognitive interview were naturally present during the think-aloud process—the absence of such codes during the think-aloud interview would provide valuable findings.

Demographic and QCAE data and analysis. Demographic data collected from students and pharmacists were summarized using descriptive statistics to illustrate the variability among participants in both groups. Quantitative comparisons between the groups were not conducted as

it was unnecessary to demonstrate the small sample sizes were sufficiently different from other another. The findings from the demographic survey are reported in chapter 4.

QCAE scoring and analysis. The QCAE instrument and scoring key was obtained with permission from the originator of the survey (Renate et al., 2011). The QCAE consists of 31 items that are mapped to either cognitive or affective empathy as described in Table 8.

Table 8

QCAE Scoring Summary

Empathy Scales	Definition	Item Numbers
Cognitive Empathy (CE)	Ability to construct a working model of the emotional states of others	Sum of PT and OS
Perspective Taking (PT)	Intuitively putting oneself in another person's shoes in order to see things from their perspective	15 – 16 – 19 – 20 – 21 – 22 – 24 – 25 – 26 – 27
Online Simulation (OS)	An effortful attempt to put oneself in another person's position by imagining what that person is feeling	1* – 3 – 4 – 5 – 6 – 18 – 28 – 30 – 31
Affective Empathy (AE)	Ability to be sensitive to and vicariously experience the feelings of others	Sum of EC, PR, and ER
Emotion Contagion (EC)	Automatic mirroring of the feelings of others	8 – 9 – 13 – 14
Proximal Responsivity (PR)	Affective response when witnessing the mood of others in a close social context	7 – 10 – 12 – 23
Peripheral Responsivity (ER)	Affective response when witnessing the mood of others in a detached social context	2* – 11 – 17* – 29*

Notes: *Items that are reverse coded (i.e. strongly agree = 1 and strongly disagree = 4)

For each item, participants are asked to evaluate their level of agreement with the statement on a 4-point Likert scale from *Strongly Disagree* (1) to *Strongly Agree* (4). Four of the items (items 1, 2, 17, and 29) are reverse coded where *Strongly Disagree* gives a score of 4 and *Strongly Agree* gives a score of 1. A participant score on the QCAE and the two subcomponents was calculated by summing the responses to items mapped to the respective subcomponents. Individuals with a higher score on the QCAE are indicative of a higher standing on the construct of empathy. The QCAE includes five subscales, however, these subscales were not mapped to SJT items included in this research as it was beyond the scope of this work and there were not a

sufficient number of participants in this pilot to analyze at the level of these subscales; therefore, the scores on the subscales are not included in this research.

The relationship between QCAE scores and other variables collected in the research study was evaluated to provide additional validity evidence and insight into the sample studied. The correlation between QCAE and SJT performance scores, for example, provides evidence to support whether this SJT was a reasonable measure of empathy; a high, positive correlation would be indicative that the instruments were measuring similar constructs. The Spearman's rank correlation coefficient was used to calculate the relationship between the QCAE and other variables (e.g. age, years of experience, etc.). This correlation coefficient was selected instead of the Pearson correlation due to the small sample size and because the Pearson correlation coefficient is not as robust in the presence of outliers, nonnormality, unequal variances, and nonlinearity (Siebert & Siebert, 2018; Siegel & Castellan, 1988). Descriptive and statistical analyses were conducted using Stata™ Version 15; correlation coefficients with a p-value < .05 were considered to identify a statistically significant relationship.

SJT performance data and analysis. Participant responses collected from this SJT were compiled for reporting purposes and were used to compare performance of the items across the different groups, the setting involved, and subcomponent of the construct evaluated. All descriptive and statistical analyses were conducted using Stata™ Version 15. Of note, the focus of this SJT response data was to provide additional validity evidence for the instrument itself and was not considered to be an exclusive component necessary to address the research questions.

SJT scoring. The use of unconventional multiple-choice test questions requires additional considerations when establishing the scoring rules. As SJTs can involve a variety of response formats, there are also a host of complementary scoring methodologies. Scoring

conventions for SJTs can be categorized as rational (i.e. pre-determined by subject matter experts) or empirical (i.e. established after large scale piloting with a sample of the testing population). Initial testing of SJTs often utilize a rational scoring convention; once large samples of test takers have completed an SJT, the empirical scoring convention is often compared for alignment. In the event the correlation of scores differs substantially between the two, subject matter experts are requested to review the scoring key for appropriateness (Bergman, Drasgow, Donovan, Henning, & Juraska, 2006). This research only included the creation of a rational key based on rankings compiled by the subject matter experts during the item development and review process. An investigation evaluating how scores differed using an empirical key was beyond the scope of this research project.

Partial credit was awarded for each item based on how much the participant differed from the rational scoring key. Table 9 provides an example of the ranking score assignment.

Table 9

Ranking SJT Item Score Matrix

Key Ranked	**Candidate Rank**				
	1	***2***	***3***	***4***	***5***
1	**4**	3	2	1	0
2	3	**4**	3	2	1
3	2	3	**4**	3	2
4	1	2	3	**4**	3
5	0	1	2	3	**4**

Each item was worth a maximum of 20 points with a possible 240 points for the total examination; four points were awarded for each response option that was ranked in the same position outlined by the key developed by the subject matter experts. For response options that differed from the key, examinees were awarded partial credit based on the distance between the correct ranking and the examinee ranking. This scoring convention was consistent with other SJTs in the health professions and has been shown to provide reliable and valid results (Patterson

et al., 2009). A Rasch partial credit approach to estimate interval differences would be ideal; however, the small sample size limits the use of robust statistical analyses using that approach (Bond & Fox, 2015).

Additional studies in the literature evaluated alternative scoring mechanisms (Bergman et al., 2006; De Lang et al., 2017). In general, scoring mechanisms are found to have impacts on reliability coefficients based on the extent to which they increase score variance using partial scoring or option weighting (Haladyna, 1990). When implementing an SJT, attention to the scoring methods and procedures must be evaluated along with alternative approaches, but changes in scoring methods should not be used to alter psychometric findings without sufficient support (Weekley, Ployhart, & Holtz, 2006). In this study, the scoring method was selected to create substantial variation in participants' total scores.

Total scores on this SJT were calculated for all participants in the research study by taking a sum of their performance on each item, which was converted to a percent score for reporting purposes based on simplicity and ease of interpretation. Total scores were intended to reflect each participant's relative standing on the construct of empathy with higher scores being indicative of a greater amount of empathy.

Psychometric analysis. The model for psychometric analysis was based on classical test theory and included the calculation of several test statistics to evaluate the test quality (AERA, APA, & NCME, 2014, *Standards* 4.9 & 4.10). Item-level and test-level statistics were calculated and summarized for the participants based on three groupings: (1) examinee experience [2 levels, student and practicing pharmacist], (2) subcomponent of the construct being evaluated [2 levels, affective and cognitive empathy], and (3) setting [2 levels, healthcare and non-healthcare].

Item-level statistics included: item difficulty (mean score and p_i = average score/points possible), standard deviation, variance, standard error, minimum score, maximum score, skewness, kurtosis, and index of discrimination (biserial correlation). Due to the rank-based response option, participant responses were also evaluated using Kendall's coefficient of concordance to determine the level of agreement among the thirty participants. The purpose of the item-level analysis was to identify if questions perform differently based on the pre-specified groupings, which could relate to observed differences found in the cognitive processes, if present.

Items were flagged as potentially problematic if they: (1) had negative item-total correlations, (2) were extremely easy ($p_i > .95$), (3) were extremely difficult ($p_i < .25$), and/or (4) had a low level of agreement of the ratings by participants (Kendall's coefficient < .6) (Luecht, 2017). Analysis of incorrect options and response patterns were limited as the item formats (i.e. ranking) create complex response options that can be difficult to interpret (Luecht, 2017). Of note, the cognitive and think-aloud interviews were used to offer insights into the observed data regarding test items. Interviews for items meeting any of these criteria were flagged but still included in the qualitative analysis; it was cautioned that these items may not assess the construct as anticipated. Special attention, however, was paid to the analysis of the interviews for these questions in the event participants identified the element that contributed to the problem identified in the statistical analysis.

Test-level statistics included: average performance, standard deviation, variance, standard error, skewness, kurtosis, minimum scores, and maximum scores. Moreover, it was necessary to evaluate the dimensionality of the test to determine the number of factors the instrument was purported to measure. There remains considerable debate regarding the dimensionality of SJTs;

it was assumed in this case that this SJT was measuring an overarching construct, therefore, it was appropriate to consider the SJT as being unidimensional (Weekley & Ployhart, 2005).

The use of a limited sample of items precludes the use of some methods (e.g., factor analysis) for assessing the internal structure of an SJT. However, a test of internal consistency (Cronbach's alpha) was calculated to evaluate how well the test items measured an intended single construct—in this case, empathy—consistently. A high Cronbach's alpha (> .8) has been recommended for SJT selection tests in the health professions (De Lang et al., 2017; Koczware et al., 2012). Other high-stakes test environments prefer higher reliability coefficients of greater than .90 or .95 (Luecht, 2017). The Spearman-Brown formula (Brown, 1910; Spearman, 1910) was applied to the results obtained to estimate the test length that would be necessary to achieve specified levels of reliability. These alpha coefficients are reported in chapter 4.

Cognitive interview coding. The first phase of the qualitative analysis focused on the cognitive interview data, which were collected to understand the response process when completing an SJT. As previously described, the response process of SJTs has remained relatively unexplored. There is minimal data to suggest an explicit structure of the cognitive process involved in SJT responses; however, models were presented in chapter 2 that outlined antecedents that affect the processes (Griffin, 2014; Lievens & Motowidlo, 2016; Ployhart, 2006) as well as models that have been used to evaluate survey response processes (Chessa & Holleman, 2007; Tourangeau, Rips, & Rasinski, 2000). These models were considered to be applicable to describing the response process of SJTs and were used to create an initial codebook for coding the transcripts from the cognitive interviews. The final codebook is provided in Appendix L.

The coding process for the cognitive interview included a calibration phase followed by three rounds of coding that were conducted by two researchers. During the calibration phase, a mock transcript was used from the pilot test of four SJT items; the two researchers coded the transcript independently according to the initial codebook and met to review discrepancies, generate example quotes for the codebook, and modify the codebook definitions as needed. The goal of the calibration phase was not to measure the level of agreement between the raters but to allow for an opportunity to align coding expectations and resolve concerns prior to the official coding process (Saldana, 2016).

Next, the cognitive interview transcript coding occurred in three rounds that involved double-coding by two researchers, auditing by a second researcher after the first research completed the coding, and independent coding by only one researcher. This process is a commonly used qualitative strategy for large data sets so that two researchers are not required to code all elements of the data (Saldana, 2016). The step-wise approach allows for frequent calibration and resolution of discrepancies without placing a large burden on the researchers while supporting consistent findings.

For the double-coding and auditing rounds, rater agreement was required to be above 80% to signify appreciable consistency between the two raters; this is consistent with expert consensus in qualitative research (Merriam & Tisdell, 2016; Saldana, 2016). Rater agreement was evaluated based on the presence of a code during each turn in the conversation (i.e. switch in the conversation from the interviewer to the participant); due to the exploratory nature of this research the frequency of codes per turn between the interview (i.e. researcher) and the participant was not delineated nor necessary. The only exception was regarding test taking strategies that were used by the participants during the study—as this was an important element

of the first research question, multiple strategies could be coded in one turn. If more than 80% agreement was not achieved in either the double-coding or auditing rounds, the subsequent rounds were to involve the same process to ensure consistency and resolve all discrepancies. For example, if the raters agreed only 75% of the time during the double-coding round then auditing would not be conducted in the subsequent round and double-coding would occur again for round two. A summary of the coding strategy and rater agreement for the cognitive and think-aloud interviews is provide in Table 10.

Table 10

Interview Coding Strategy and Rater Agreement

Interview Type	Coding Strategy	Items	Average Rater Agreement
Cognitive Interview			
Round 1	Double-Code	CH1, CN1, AH1, AN1	80.2%
Round 2	Audit	CH2, CN2, AH2, AN2	97.7%
Round 3	Independent	CH3, CN3, AH3, AN3, Concluding Questions	---
Think-Aloud Interview			
Round 1	Double-Code	S02, S08, S09, S14, S15 P01, P04, P06, P11, P13	87.5%
Round 2	Audit	S03, S06, S07, S10, S13 P02, P05, P08, P09, P10	94.9%
Round 3	Independent	S01, S04, S05, S11, S12 P03, P07, P12, P14, P15	----

In the first round of coding the cognitive interviews, both researchers were required to independently code four SJT items: CH1, CN1, AH1, and AN1. For the first item (CH1), the researchers coded the transcripts based on the initial codebook that was established using evidence in the literature about factors that may influence the response process. Researchers were also permitted to inductively code in which they could label segments of text as "other" if they identified what they perceived to be an emerging code that was not identified in the initial codebook. The researchers met after coding the CH1 transcript to discuss discrepancies and modifications to the codebook.

During this review session of the CH1 transcript codes, the researchers identified six new codes that were added to the codebook. Two were related to the SJT framework; this included participants suggesting a lack of experience or the description of an experience of knowledge that could not be reliably identified as relating to either a healthcare or non-healthcare setting. In addition, four codes were related to the response process framework. These codes included: assumptions examinees used to answer the questions (e.g. assuming the type of tone portrayed by a character in the scenario or assuming constraints of the situation), feelings about the test (e.g. whether an item was difficult), strategies used to answer test questions (e.g. identifying the best and worst responses first), and pertinent contextual elements (e.g. identifying their response would change based on the relationship with the individual in the scenario). These six additional codes were integrated into the final codebook and additional details about these codes and samples are provided in Appendix L.

The two researchers independently coded transcripts using the revised codebook for CH1 again, in addition to coding the transcripts for CN1. The researchers engaged in inductive coding during this coding process as well, however, no new codes emerged. The researchers met to review the coding for CH1 and CN1 and resolved all discrepancies. The same process was completed for the AH1 and AN1 transcripts. The average rater agreement prior to resolving discrepancies for the double-coding process in round one across the four items was 80.2%, which allowed the researchers to use an audit approach for round two.

During round two, the primary researcher independently coded items CH2, CN2, AH2, and AN2 using the final codebook. The second researcher then independently reviewed the coded transcripts. The second researcher was required to note if he or she agreed with the code provided by the primary researcher and to include any coding they believe was missed by the

primary researcher. The two researchers then met to resolve discrepancies. The average rater agreement during the second round of coding was 97.7%.

The final round of coding was conducted independently by the primary researcher and included SJT items CH3, CN3, AH3, and AN3 as well as coding of the general questions participants were asked at the end of the cognitive interview. For this round, the second researcher was not included as the previous round of coding met the criterion of greater than 80% agreement. Coding of the cognitive interviews concluded with a set of the final transcripts that included the agreed upon codes by the researchers. This final set of coded cognitive interview transcripts was used in the subsequent data analysis.

Think-aloud interview coding. The think-aloud interview analysis was intended to confirm cognitive models that describe the problem-solving process used during SJT completion (Leighton, 2017). The think-aloud interviews were coded using the same process used to code the cognitive interviews described in the previous section and outlined in Table 10. During the coding of the think-aloud interviews there were no new codes added to the codebook provided in Appendix L. Of note, the coding of the think-aloud interview was conducted by the primary researcher and a third researcher who was not involved in the cognitive interview coding process. The goal was to include a different perspective during this process to avoid potential bias that could occur after reviewing the cognitive interviews. Rater agreement exceeded the 80% threshold with 87.5% agreement during the double-coding round and 94.9% agreement during the auditing round. Coding of the think-aloud interviews concluded with a set of the final transcripts that included the agreed upon codes by the researchers. This final set of coded think-aloud interview transcripts was used in the subsequent data analysis

Cognitive and think-aloud interview data analysis. The final coded transcripts served as the main data sources to address the proposed research questions. Overall, the presence and distribution of codes identified in the participant interviews was pivotal in answering the research questions. The coded transcripts were reviewed for the prevalence and context of the utterances shared by the participants. Themes were identified by looking for patterns and relationships between the present codes in addition to what participants shared about their experience. These themes and conclusions were then reviewed by other research team members to determine if the findings were sufficiently supported based on the evidence.

Addressing the Research Questions

The following section articulates how the qualitative data were used to address each of the research questions presented in this study. Each of the research questions is provided for reference with a brief discussion of how the analysis informed the results presented in chapter 4. As outlined in Table 7, the think-aloud interviews primarily addressed the first research question whereas the cognitive interviews were intended to address all four research questions. There were few instances in which the think-aloud interviews contributed substantially to the remaining questions as participants did not make explicit statements pertaining to these questions.

RQ1: What factors and strategies are involved in the cognitive processes when examinees respond to SJT items? To answer this question, the codes from the cognitive interviews and think-aloud interviews transcripts were reviewed to identify common patterns and prevalent themes that emerged based on the content of the utterances shared by examinees during and about an SJT. The presence of codes was described across SJT items to determine which features were most salient compared to others and how those codes supported the frameworks outlined in chapter 2.

The cognitive response process, for example, was suspected to include elements of comprehension, retrieval, judgment, and response selection. In addition, there are antecedents related to experience, knowledge, and personal factors that were expected to influence the decision-making process. The analysis was intended to identify which features were and were not consistently present in the cognitive and think-aloud interviews. The high prevalence of certain codes was suggestive that these were essential elements in SJT response processes of examinees, whereas codes that were not present were not considered relevant in the response process.

The distribution of these codes and themes across the cognitive interviews versus the think-aloud interviews was also described and explored. The two interview types had different purposes; the think-aloud was intended to describe what occurs naturally when the examinee completes an SJT whereas the cognitive interview included specific questions to probe the participant about the research questions explicitly. Observed differences in the prevalence of codes and themes was indicative of which cognitive processes were engaged more readily by the examinee. In other words, if a code or theme was only prevalent in the cognitive interview and not in the think-aloud interview, it suggested that those components may not inherently be used by the examinee and only identifiable when asked; therefore, these features were not considered to be as relevant in SJT response processes.

RQ2: What is the role of job-specific experiences (i.e. student or experienced clinicians) in the response process to SJT items? To answer this question, examinees were asked during the cognitive interview to identify how they thought about prior experiences when addressing the selected SJT question. The responses were coded, and the results were summarized to describe

which types of experiences were most prevalent and how the experiences differ between the groups of examinees (i.e. whether they were a student or experienced pharmacist).

For example, if both students and experienced pharmacists consistently failed to recall job-specific experiences, it would be indicative that the job-specific experiences were not as critical as anticipated in the answering some SJT items. Conversely, if experienced pharmacists tended to discuss job-specific experiences more often than student pharmacists, this would indicate job experiences were an important element in making more informed decisions on SJTs.

RQ3: What is the role of the setting presented in SJT items in the response process (i.e. the influence of a healthcare or non-healthcare specific setting)? To answer this research question, examinees were asked during the cognitive interview to describe how the setting was important in answering the question. Coded responses were summarized and compared for the items that were designed to be healthcare-specific and those that were not. Half of the test was designed with questions that were not specific to the health professions to explore if examinees recognized this difference and the types of knowledge or experiences that influenced their response in either setting.

The distribution of the codes and themes across these two groups of items was investigated to determine if some features were frequently explicated more often by examinees based on the setting of the item. In general, the extent to which examinees refer to the setting when responding to the item signified an important indicator about the role of the setting. If examinees rarely discussed features of the question being in a healthcare setting or if they did not mention it as an influence in their decision-making process, it would indicate the setting did not contribute significantly to that test question.

RQ4: What is the role of the ability to identify the construct being evaluated (i.e. empathy) in the response process to SJT items? To answer this question, examinees were explicitly asked during the cognitive interview what they believed the question was asking them to or what they feel was being assessed. The responses were summarized to identify how often empathy was identified as the construct being tested and how determining this feature was related to performance on the item. In addition, all other constructs that were suspected by participants when asked this question were reported at the item level to determine if a construct other than empathy was being assessed. If many of the participants reported empathy as the construct being assessed, it would provide validity evidence supporting that this SJT was measuring the desired construct. If multiple examinees reported a different construct being measured for a specific item, the item and examinee utterances were reviewed further to identify why they suspected that construct and offer strategies for design modifications in the future.

Summary

The purpose of this research was to explore the response processes participants used to address scenarios presented to them during SJTs. Two groups of participants with differing levels of experience were asked to complete a 12-item SJT intended to assess two subcomponents of empathy: affective empathy and cognitive empathy. In addition, this SJT included questions that varied based on their setting (i.e. healthcare and non-healthcare) to determine how these differences may influence the response process. Participants engaged in a think-aloud interviewed while they completed the SJT followed by a cognitive interview with the researcher who asked questions to better understand the cognitive processes used when completing SJT items. The transcripts of the interviews were coded and analyzed to determine the prevalence of major codes and themes that addressed the presented research questions.

Chapter 4: Results

This chapter presents the results of administering an SJT intended to measure empathy to students and practicing pharmacists. The provided results are intended to address significant gaps in the literature regarding the response process involved in an SJT. The primary goal was to provide evidence of the salient factors that may influence the response process and to describe how these factors align with the current—albeit limited—understanding of SJT response processes. More specifically, the results were aimed to investigate significant questions about SJT response processes, including understanding the role of participants' experiences recalled during the testing process and understanding the influence of the item setting on response selections. Lastly, the results contributed to a growing interest in SJT research, which is to better describe the role of the participant's ability to identify the construct being assessed and investigating that relationship with SJT performance. Cognitive and think-aloud interviews were used to generate data concerning the response process, which were analyzed using quantitative and qualitative methodologies in this comprehensive and exploratory approach. Overall, the results make significant contributions to the emerging body of validity evidence regarding the interpretation of SJT scores.

The chapter begins with a summary of the development process for this SJT in which the instrumentation development process followed an evidence-centered approach to target the desired construct—empathy. The chapter continues with a description of the study participants, a summary of their characteristics, and the results of administering the *Questionnaire of Cognitive*

and Affective Empathy (QCAE). Next, a psychometric analysis of this SJT is presented based on the participants' responses to this SJT. The chapter then provides results for the research questions described previously based on the cognitive interviews and think-aloud interviews.

Summary of SJT Instrumentation Development

A 12-item SJT was developed using evidence-centered design principles to optimally create an instrument that targeted one construct of interest (i.e., empathy). A panel of 11 subject matter experts—including practicing pharmacists and pharmacy faculty—with an average of 13.5 years of experience across multiple specialties were recruited to assist with this SJT development. These individuals participated in either the item design phase or the item review process.

During the item design phase, seven of the experts created 24 items designed to measure two subcomponents of empathy (i.e., affective or cognitive) in various settings (i.e., healthcare or non-healthcare). During the item review process, four experts independently evaluated each item on three criteria: (1) how well the item measured empathy, (2) the subcomponent of empathy assessed, and (3) the type of setting used in the item. Each expert was also required to rank the response options to determine the level of agreement in the final key for the item, which served as an additional evaluation criterion. Evaluations of the items were aggregated to create a series of indices used to judge each item against pre-determined criteria to determine if the item should be included in the final instrument; this approach is described extensively in chapter 3. Of the 24 items, 12 were included in the final SJT administered to participants. SJT items were evenly distributed in the number that measured the subcomponents of empathy (i.e., 6 items measured either cognitive or affective empathy) and the setting (i.e., 6 items were either in a healthcare or non-healthcare setting).

The final 12-item SJT was converted to a paper test that was provided to study participants. The test order was randomized for each participant to minimize order effects. Participants were provided an alphanumeric identifier to designate if they were a student participant (indicated by the label "S" followed by a number from 1 to 15) or a pharmacist participant (indicated by the label "P" followed by a number from 1 to 15). References to specific participants in the remainder of this chapter use these alphanumeric identifiers to ensure anonymity. The subsequent sections of this chapter present the data collected exclusively during the administration of this SJT and from interviews of study participants, all of whom were not included in the design of the instrument.

Sample Characteristics from SJT Administration

A total of 30 participants consented to participate in the study; 15 participants were students and 15 participants were licensed pharmacists with at least five years of experience. The goal was to include individuals with varied backgrounds, which was successfully achieved. Data presented are grouped by participant type as this was the first research study that evaluated if there were significant differences in performance based on the level of experience of examinees (i.e., student compared to experienced pharmacists). Table 11 provides a detailed summary of the characteristics of the sample.

Student characteristics. The student group was predominantly female (n = 11, 73.3%) with a median age of 24 years (range 22-45 years). Most of the students were entering their third or fourth year of pharmacy school (n = 11, 73%), which means they have some experience working in a pharmacy practice setting through rotation experiences. In addition, 13 of the students (87%) indicated working in a healthcare-related field outside of their coursework. Eight of the students (53%) reported working in a non-healthcare human services field with a one year

of experience being the median (range 0-10 years). Eighty percent (n = 12) of students reported having training related to empathy; they most often cited coursework or classroom discussions regarding mental health and working with patients.

Table 11

Participant Characteristics by Participant Type (N = 30; median and range unless noted)

	Students (n = 15)	Pharmacists (n = 15)
Male, n (%)	4 (27)	2 (13)
Age	24 (22-45)	36 (29–51)
Anticipated graduation year, n (%)		
Class of 2019	4 (27)	***
Class of 2020	7 (40)	***
Class of 2021	4 (27)	***
Education and training, n (%)		
Bachelor of Science Degree	15 (100)	***
Doctor of Pharmacy	***	15 (100)
Residency (e.g. PGY1 and/or PGY2)	***	13 (87)
Fellowship or post-doc	***	1 (7)
Advanced degree (e.g. MPH, MBA, PhD)	***	3 (20)
Board certification (e.g. BCPS, BCOP)	***	11 (73)
Practice Location, n (%)		
University hospital A	***	11 (73)
University hospital B	***	4 (27)
Practice area, n (%)		
Academia	***	2 (13)
Administration	***	1 (7)
Ambulatory care	***	1 (7)
Cardiology / pulmonology	***	2 (13)
Critical care / emergency medicine	***	3 (20)
General medicine	***	2 (13)
Infectious diseases	***	2 (13)
Psychiatry	***	1 (7)
Surgery	***	1 (7)
Work experience in a healthcare-related field, n (%)	13 (87)	15 (100)
Years licensed as a pharmacist	***	8 (6–23)
Years with a health professions faculty appointment	***	5 (0–20)
Average number of hours working in a healthcare setting per week	5 (0–40)	40 (0–55)
Average number of patients interacting with per week	3 (0–75)	20 (0–100)
Average number of students interacting with per week	***	2 (0–8)
Average number of non-pharmacist healthcare providers interacting with per week	2 (0–10)	10 (2–35)
Work experience in a nonhealthcare-related human services field, n (%)	8 (53)	11 (73)
Years of work experience in a nonhealthcare-related human services field	1 (0–10)	4 (0–10)
Experience taking care of a terminally ill family member or individual, n (%)	1 (7)	4 (27)
Questionnaire of Cognitive and Affective Empathy (QCAE) Score	93 (79–103)	90 (85–105)
Cognitive Empathy (CE) Score	57 (46–67)	58 (50–68)
Affective Empathy (AE) Score	37 (27–42)	34 (29–39)
Self-reported training related to empathy, n (%)	12 (80)	5 (33)

Pharmacist characteristics. The pharmacist group was also predominantly female (n = 13, 86.6%) with a median age of 36 (range 29-51 years). Participating pharmacists worked in various practice areas, but all were employed in a university hospital setting. A majority of the pharmacists completed a residency (n = 13, 87%) and were board certified (n = 11, 73%); this indicates that these individuals have extensive training in specialty areas and providing advanced patient care. Eleven of the pharmacists (73%) reported working in a nonhealthcare human services field with a median of 4 years of experience (range 0-10 years). Only 33% (n = 5) of pharmacists reported having training related to empathy; participants frequently cited exposure to material related to emotional intelligence or service recovery training specific to their institution.

QCAE results. This SJT included in this study had not been rigorously tested prior to its use with large samples, therefore, an additional instrument to measure participant empathy was included. All 30 participants completed the QCAE, which provides a self-reported measure of cognitive and affective empathy (Renate et al., 2011). Scores on the QCAE can range from 31 to 124; the score is the sum of the cognitive empathy (CE) sub-score (range of 19 to 76) and the affective empathy (AE) sub-score (range of 12 to 48). The mean score on the QCAE was 91.8 (SD 6.1) and total scores ranged from 79 to 105. The mean CE and AE sub-scores were 57.1 (SD 5.4) and 34.7 (SD 3.8), respectively. Results of a Mann-Whitney test suggested non-significant differences ($p > .05$) between the median QCAE, CE, and AE scores for participants in the student and pharmacist groups. This finding implies that the pharmacy students and licensed pharmacists included in this study scored similarly and that their standings on measures of cognitive, affective, and overall empathy were not substantially different. In other words, these samples did not differ significantly in their levels of empathy.

Relationship of the QCAE to other variables. The correlation of QCAE scores and sample characteristics were explored to determine if there were significant relationships that could be pertinent in understanding SJT performance as it pertains to empathy. Spearman's rank correlation coefficients were calculated to describe the relationship between QCAE and continuous variables such as age, years licensed, years working in healthcare settings, and overall SJT performance; the results are presented in Table 12a. Point biserial correlations were calculated to describe the relationship between QCAE scores and dichotomous variables such as whether the participant was a pharmacist, whether the individual had healthcare or service-related work experiences, and whether the individual reported previous empathy training; the results are presented in Table 12b.

There were few statistically significant relationships between the QCAE and other variables that provided meaningful findings. Relationships of note include that the AE sub-score was positively correlated with being a female (r_{bis} = .41, p < .05); this finding was consistent with previous findings that female gender is often positively correlated with higher levels of empathy, especially affective empathy (Renate et al., 2011). In addition, self-reported empathy training had different relationships to QCAE scores for students and pharmacists; for students, empathy training was negatively correlated with QCAE scores (r_{pbis} = –.51, p = .06) whereas for pharmacists, training was positively correlated with QCAE scores (r_{pbis} = .47, p = .08). This finding suggests that student-specific training may differ from pharmacist-specific training about empathy; however, a majority of students (80%) reported having training about empathy. Students often referenced classroom discussions as the source of this training, whereas pharmacists referenced specific on-the-job training. Therefore, this finding should be interpreted

with caution as there may be significant variations in perceptions of what qualifies as empathy training and it does not account for differences in high-quality training.

Table 12a

Spearman's Rank Correlation Coefficients of the QCAE to Other Variables

	All participants (n = 30)			Students (n = 15)			Pharmacists (n = 15)		
	QCAE	*CE*	*AE*	*QCAE*	*CE*	*AE*	*QCAE*	*CE*	*AE*
QCAE Total Score	***	**.67**‡	**.52**‡	***	**.71**‡	**.55***	***	**.76**‡	.37
CE Score	**.67**‡	***	–.20	**.71**‡	***	–.13	**.76**‡	***	–.16
AE Score	**.52**‡	–.20	***	**.55***	-.13	***	.37	–.16	***
Age	–.25	–.21	–.08	-.33	**–.63***	.29	–.39	–.12	–.34
Years licensed	***	***	***	***	***	***	–.38	–.12	–.23
Years as faculty	***	***	***	***	***	***	–.17	.11	–.44
Weekly HC hours	.09	.01	.14	.35	–.07	**.54***	.26	.10	.35
Number of patients	.04	.05	.03	.33	.03	.33	–.30	.01	–.28
Number of students	***	***	***	***	***	***	–.19	.03	–.15
Number of HC providers	.19	.16	.02	.32	.14	.21	.38	.14	.42
Non-HC work experience	.08	–.04	.18	.27	.03	.20	–.21	–.21	–.05
Years in non-HC	.08	–.08	.76	.25	.11	.23	–.18	–.38	–.01
SJT performance	**.34**^	.00	**.32**^	.35	–.06	**.65**‡	.31	.03	.07

Notes: QCAE = total QCAE score, CE = cognitive empathy, AE = affective empathy, HC = healthcare
^ $p < .10$, * $p < .05$, ‡ $p < .01$

Table 12b

Point Biserial Correlations of the QCAE to Other Variables

	All participants (n = 30)			Students (n = 15)			Pharmacists (n = 15)		
	QCAE	*CE*	*AE*	*QCAE*	*CE*	*AE*	*QCAE*	*CE*	*AE*
Female	.25	–.01	**.41***	.33	–.11	**.64***	.13	.14	.01
Pharmacist	–.03	.09	–.18	***	***	***	***	***	***
University Hospital A	***	***	***	***	***	***	.20	.04	.34
HC work experience	***	***	***	.08	–.06	.20	***	***	***
Non-HC work experience	.08	–.04	.18	.27	.03	.20	–.21	–.21	–.05
Care for terminally ill	.18	.25	–.06	.15	.02	.21	.27	**.47**^	–.28
Empathy training	–.04	–.14	.14	**–.51**^	**–.57***	.01	**.47**^	**.45**^	.16

Notes: QCAE = total QCAE score, CE = cognitive empathy, AE = affective empathy, HC = healthcare
^ $p < .10$, * $p < .05$, ‡ $p < .01$

The most critical relationship investigated was the correlation between QCAE score and SJT performance; this relationship was essential as it provided foundational validity evidence for the administered SJT in that a positive correlation would suggest this SJT was successful at targeting empathy. The results showed a moderate, positive correlation ($r_s = .34$, $p = .07$) between the two variables, which suggested this SJT and the QCAE were measuring similar

constructs (i.e., empathy). This finding supported moving on to the psychometric analyses of SJT performance data to describe how well participants performed on this SJT, to describe how performance differed based on item characteristics and participant type, and to identify problematic items that may need to be excluded from subsequent analyses with regards to cognitive and think-aloud interviews.

Psychometric Properties of this SJT

Each question was scored based on the empirical key created by the subject matter experts during the item development phase. Participants were awarded partial credit based on how well their ranking of the response options matched the ranking determined to be more appropriate by the subject matter experts. Each question was worth 20 points with 240 points possible on the entire test. Table 13 includes a summary of the item-level and test-level statistics for this SJT, which are based on participant performance data provided in Appendix M.

Table 13

SJT Item Psychometrics Based on All Participant Responses (N = 30)

	M	*SD*	*r*	*Min*	*Max*	*W*
CH1	15.6	3.1	.23	10	20	.66
CH2	15.1	2.7	.06	12	20	.65
CH3	17.1	2.5	.14	12	20	.81
CN1	13.8	2.8	.26	8	20	.76
CN2	15.0	2.6	.63	10	18	.54
CN3	13.9	3.3	.38	8	20	.68
AH1	15.7	3.6	.49	8	20	.69
AH2	13.2	3.0	.30	8	20	.56
AH3	17.3	1.9	.38	14	20	.85
AN1	14.7	3.4	.40	8	20	.77
AN2	15.5	2.6	.35	10	20	.66
AN3	13.7	3.0	.08	8	20	.39
TOTAL TEST	180.6	11.8	***	142	200	***

Notes: M = mean, SD = standard deviation, r = discrimination index (Pearson's r), Min = minimum score, Max = maximum, score, W = Kendall's coefficient of concordance

The average score on this SJT across the 30 participants was 180.6 (75.3%) with a range of 142 (59.2%) to 200 (83.3%); the standard deviation in test scores was 11.8. Data from the overall test performance exhibited negative skewness (–.93) and showed substantial positive

kurtosis (5.1). Performance on all the items was positively correlated with total score on this SJT based on the Pearson's correlation coefficients; the most discriminating items were CN2, AH1, and AN1. Kendall's coefficient of concordance was used as an additional indicator of agreement in the participant ranking of response options and values less than .6 are indicative of poor agreement. Three items (CN2, AH2, and AN3) had coefficients of concordance below this value (.54, .56, and .39, respectively), which suggests there may be disagreement in the rankings provided by the participants (i.e. greater variability in the response patterns for these items).

The psychometric properties of this SJT were also evaluated based on pertinent variables of interest in this research including: the item setting (e.g., healthcare or non-healthcare), the empathy component assessed in the item (e.g., affective or cognitive), participant-type (e.g., student or pharmacist), and item groups based collectively on the setting and empathy component assessed. The mean, standard deviation, minimum, and maximum values according to these classifications are provided in Table 14. Overall, there was no evidence to suggest there were significant differences in SJT performance as it pertains to any of these classifications.

Cronbach's alpha was calculated to evaluate the internal consistency of the items that make up this SJT; the expectation was that all items were measuring a unified construct (e.g. empathy), therefore, a high alpha would indicate the instrument was consistently targeting one construct. The Cronbach's alpha was equal to .30 based on the full SJT (i.e., 12 items). Additionally, Cronbach's alpha was also calculated for items related to cognitive empathy (α = .06) and affective empathy (α = .22). Of note, the low observed alpha values are likely attributed to the small numbers of items, small number of participants tested, highly variable inter-item correlations, and homogeneity of the participant sample. Table 15 provides the correlation matrix of SJT items for reference based on all participant responses, which shows significant variation

in the items that are positively and negatively correlated with one another. There were six item pairs with statistically significant Pearson's correlation coefficients; however, most of these were items that measured similar subcomponents of empathy or were located within similar settings.

Table 14

SJT Item Psychometrics by Item and Participant Classifications Based on All Participant Responses (N = 30)

	n_i	*M*	*SD*	*Min*	*Max*	*Alpha*
Setting						
Healthcare	6	15.6	3.1	8	20	.25
Non-healthcare	6	14.5	3.0	8	20	.52
Empathy Component						
Affective	6	15.0	3.2	8	20	.22
Cognitive	6	15.1	3.0	8	20	.06
Participant						
Student	15	14.9	3.1	8	20	***
Pharmacist	15	15.2	3.1	8	20	***
Item Grouping						
CH	3	15.9	2.9	10	20	.16
CN	3	14.2	2.9	8	20	.52
AH	3	15.4	3.3	8	20	.25
AN	3	14.7	3.1	8	20	.22
TOTAL TEST	12	180.6	11.8	142	200	.30

Notes: n_i = number of items in item category or participant type; M = mean, SD = standard deviation, Min = minimum, Max = maximum, Alpha = Cronbach's alpha

Overall, the results of the psychometric analyses suggested that the SJT developed for this study was capable of providing a reasonable estimate of participants' empathy given the constraints of a small sample size and brief scale (i.e., 12 total items). Although some items did not perform optimally (e.g. AH3) and there was limited evidence that this SJT targeted a unidimensional construct, there was no indication that data pertaining to any item should be excluded from further analysis. The following sections of this chapter describe results from the cognitive and think-aloud interviews to explicitly address the four research questions.

Table 15

Correlation Matrix of SJT Items (N = 30)

	CH1	*CH2*	*CH3*	*CN1*	*CN2*	*CN3*	*AH1*	*AH2*	*AH3*	*AN1*	*AN2*
CH1	***										
CH2	.24	***									
CH3	–.01	–.01	***								
CN1	–.13	**–.49**‡	–.09	***							
CN2	–.02	–.01	–.11	.29	***						
CN3	.05	.04	–.15	.14	**.49**‡	***					
AH1	.19	–.06	.12	–.09	.20	.08	***				
AH2	–.12	–.14	.06	.06	.18	–.09	.06	***			
AH3	**.39***	.00	–.03	–.26	.00	.27	**.48**‡	.06	***		
AN1	.05	.03	.04	–.11	.26	.10	.16	.11	.16	***	
AN2	–.22	–.03	–.17	.07	.28	–.05	.08	**.36***	–.11	.15	***
AN3	–.28	–.07	–.16	**.37***	.05	–.02	–.23	–.18	–.27	–.08	.16

Notes: * $p < .05$, ‡ $p < .01$

RQ1: Factors and Strategies Involved in the SJT Response Process

The first research question was: *"What factors and strategies are involved in the cognitive processes when examinees respond to SJT items?"* Specifically, the goal was to determine the extent to which previously identified features of SJT response processes (see Griffin, 2014; Lievens & Motowidlo, 2016; Ployhart, 2006; Tourangeau, Rips, & Rasinksi, 2000) were evident during cognitive and think-aloud interviews. This question was a significant component of the research as previous studies about SJTs do not offer a comprehensive framework or substantial investigation into the response process. In addition, the predominantly qualitative approach could identify other factors or strategies not previously documented in SJT research.

The following portions of this section summarize the data analysis process for the first research question, describe the most and least prominent codes present based on various classifications (i.e., interview-type, participant-type, setting, and, empathy component), present an integrated model of SJT response processes that adds new factors, and include participant reflections about SJTs (e.g., what made items easier, harder, etc.).

Data analysis summary. To answer the first research question, both the cognitive interviews and think-aloud interviews served as essential data sources. During the cognitive interviews for each SJT item, 10 students and 10 pharmacists were asked about how they arrived at the final ranking of the response options for eight of the 12 SJT items. In addition, participants were asked in the cognitive interview to describe what made their decisions on each item easier or more difficult. At the conclusion of the cognitive interviews, participants were also asked broadly about what factors they believe contributed to their performance on this SJT as well as the factors that made the entire test easier or more difficult.

Transcripts from the cognitive and think-aloud interviews were coded according to the codebook (Appendix L), which included factors expected to be present based on existing frameworks about SJTs and test response process (Griffin, 2014; Lievens & Motowidlo, 2016; Ployhart, 2006; Tourangeau, Rips, & Rasinksi, 2000). The codebook was modified during the initial review of transcript data to include additional sub-codes pertaining to factors not previously documented in the frameworks. Two sub-codes, *objectives* and *assumptions*, were added to better describe the comprehension process and three sub-codes, *perceptions*, *feelings about the test*, and *context*, were added to better classify judgments during the response process. Lastly, the *strategies* sub-code was added to describe general approaches participants used to select final answers. Definitions and examples of these sub-codes are included in the final codebook (Appendix L); a discussion of how these sub-codes pertain to the proposed model is provided later in this section.

The frequency of codes across cognitive and think-aloud interviewers was compared to determine which factors of the framework were most prevalent and whether this differed according to the item and participant classifications being studied. Least prevalent codes were

also identified as this research was the first to offer a comprehensive analysis into SJT response processes; therefore, it was pertinent to identify if codes expected to be present according to previous research were observed in the cognitive and think aloud interviews. Coded segments were aggregated and quantified across items and participants to investigate if there was a pattern regarding the factors involved in SJT response processes. Coded segments within sub-codes were categorized based on common patterns to better describe pertinent features of the response process, especially features that were not previously identified in the literature. A heat map, provided in Appendix N, was also created as a strategy to visualize patterns across items and participant types. Of note, sub-codes were not included in the frequency counts of the overarching code to avoid duplicative frequencies that would artificially inflate the presence of the overarching code.

Prevalence and distribution of codes. In summary, there were a total of 7,252 coded statements distributed across 30 cognitive and think-aloud interviews. Approximately 18.4% of all coded segments pertained to judgments, which included making decisions or value-statements that were generated by integrating memories, knowledge, experiences, and personal factors. The other most prevalent codes included: comprehension (13.4%), retrieval (8.3%), emotional intelligence (7.3%), and objectives (7.0%). The least prevalent codes across all interviews included: general knowledge (0.2%), ability (0.3%), nondescript experiences (0.6%), affective empathy (0.8%), and impression management (0.9%). These data suggest there are definitive salient components of the presented frameworks (i.e., judgments, retrieval, emotional intelligence, etc.), however, not all components (i.e., general knowledge, ability, etc.) may be as critical in the response process or they may not be overtly described by participants using these methodologies. At this exploratory phase of research, the low prevalence of a code was not

considered to be sufficient evidence to completely remove it from the proposed response process model.

Of note, the prominence of codes differed depending on the source of the data (i.e., cognitive interviews compared to think-aloud interviews), As shown in Table 16, codes that were consistently prevalent regardless of interview type included judgments and emotional intelligence. Cognitive interviews were more likely to include references to retrieval, response selection, and perceptions that influenced their responses; whereas, think-aloud interviews included more references about the task objective, the context of the item, and feelings about the test. With regards to the least prevalent codes, there were few differences based on the interview type. Both interview types rarely included references to general knowledge and ability. The cognitive interview differed in that it did not include references to nondescript experiences and impression management, whereas the think-aloud interview did not include references to a lack of experience and perceptions that influence response selection.

In summary, there is evidence to support that many of the features of SJT response processes described in the literature were present based on findings in the cognitive and think-aloud interviews. Due to the structured approach of the cognitive interview, it is possible some codes were more prevalent because questions were specifically asked of participants during that type of interview. Conversely, the distribution of codes throughout the think-aloud interviews are assumed to be more indicative of the natural response process. Although some codes were not highly prevalent in either interview, there is not sufficient evidence to suggest these factors are insignificant in SJT response processes.

Table 16

Most and Least Prevalent SJT Response Process Codes Based on Interview Type

	Most Prevalent Codes		Least Prevalent Codes	
	Code	*% Total Codes*	*Code*	*% Total Codes*
Cognitive	Judgment	17.4	**Nondescript Experience**	.2
Interview	Comprehension	12.9	General Knowledge	.2
	Retrieval	10.7	Ability	.3
	Response Selection	8.9	**Impression Management**	.7
	Emotional Intelligence	8.3	**Affective Empathy**	.8
Think-Aloud	Comprehension	17.7	General Knowledge	.1
Interview	Judgement	14.9	**Lack of Experience**	.1
	Objective	14.6	**Perceptions**	.1
	Context	13.0	Ability	.2
	Feelings about the Test	7.9	**General Experience**	.3
Total	Judgment	18.4	General Knowledge	.2
	Comprehension	13.4	Ability	.3
	Retrieval	8.3	Nondescript Experience	.6
	Emotional Intelligence	7.3	Affective Empathy	.8
	Objective	7.0	Impression Management	.9

Notes: **Bold** = difference between the interview types (i.e. cognitive compared to think-aloud interview)

Distribution of codes by item classification and participant type. The distribution of codes was also examined with respect to three classifications: item setting (i.e. healthcare and non-healthcare), item empathy component (i.e., affective and cognitive), and participant type (i.e., student and pharmacist). Table 17a includes a summary of the most prevalent codes according to interview type and the three classifications. To readily identify differences, an asterisk was used to indicate differences within the same interview type, whereas, a double-cross was used to indicate differences between the interview types.

In general, there were minimal differences in the prevalence of codes with regards to the item classification and participant types. For example, in the cognitive interviews the only difference in the most prevalent codes based on setting was that the healthcare questions had more references to objectives whereas the non-healthcare questions had more references to the ability to identify the construct. Differences were more common between cognitive and think-aloud interviews, however, interpretation of these results must be done carefully as each methodology is designed to elicit certain responses from participants that may contribute to

observed differences. Of note, reference to response selection, task objectives, and perceptions that influenced response choices were more common in think-aloud interviews than in cognitive interviews, which was consistent across item classifications and participant types.

Table 17a

Most Prevalent Codes During Cognitive and Think-Aloud Interviews Organized by Item Classification and Participant Type

	Cognitive Interview			Think-Aloud Interview			Total		
Code	*Set*	*Emp*	*Part*	*Set*	*Emp*	*Part*	*Set*	*Emp*	*Part*
Retrieval	H, N‡	A‡, C‡	S‡, P‡	H*			H, N	A, C	S, P
Comprehension	H‡, N‡	A‡, C	S‡, P‡		C*		H, N	A, C	S, P
Judgment	H, N	A, C	S, P	H, N	A, C	S, P	H, N	A, C	S, P
Response Selection				H‡, N‡	A‡, C‡	S‡, P‡	H, N	A, C	S, P
Emotional Intelligence	H‡, N	A, C	S, P	N*	A, C	S, P	H, N	A, C	S, P
Context	H‡, N‡	A‡, C‡	S‡, P‡				N*		P*
Objective	H*		S*	H, N‡	A‡, C‡	S, P‡	H*	A, C	S*
Perceptions				H‡, N‡	A‡, C‡	S‡, P‡			
Ability to Identify Construct	N*	A‡, C‡	P*						

Notes: Set = setting, H = healthcare setting, N = non-healthcare setting,
Emp = empathy subcomponent, A = affective empathy, C = cognitive empathy
Part = participant, S = student, P = pharmacist
*difference within same interview type (e.g., difference between setting, empathy, or participant)
‡difference between interview types (e.g., cognitive compare to think-aloud interview)

A similar process was conducted for the least prevalent codes; Table 17b includes a summary of the least prevalent codes according to interview type and the three classifications. The goal of identifying the least prevalent codes was to determine if there were features of SJT response processes previously suspected to be pertinent that were not observed in the cognitive or think aloud interviews. The lack of a code would suggest that the feature may not be as critical as reported in SJT research. To readily identify differences, an asterisk was used to indicate differences within the same interview type, whereas, a double-cross was used to indicate differences between the interview types. In this case, there were substantial differences between classifications and interview type. These results, however, were not considered to be a significant

finding as the frequencies of these codes were very small (i.e., less than 10 per classification), which means small variations could contribute to significant differences.

Table 17b

Least Prevalent Codes During Cognitive and Think-Aloud Interviews Organized by Item Classification and Participant Type

	Cognitive Interview			Think-Aloud Interview			Total		
Code	*Set*	*Emp*	*Part*	*Set*	*Emp*	*Part*	*Set*	*Emp*	*Part*
General Knowledge	H, N	A, C	S, P	H, N	A, C	S, P	H, N	A, C	S, P
General Experience				H*					
Job-Specific Experience				N*‡		P*‡			
Nondescript Experience	N*‡	A*	S*‡		A*		N*	A*	S, P
Lack of Experience				H‡, N‡	A‡, C‡	S‡, P‡			
Impression Management	H, N‡	A, C‡	S, P	H*	A*	S, P	H, N	A, C	S, P
Ability	H, N	A‡, C	S, P	H, N	C*	S, P	H, N	A, C	S, P
Affective Empathy	H*‡	C*	P*‡		C*		H*	C*	
Ability to Identify Construct				N*‡	A*‡	S*‡			

Notes: Set = setting, H = healthcare setting, N = non-healthcare setting,
Emp = empathy subcomponent, A = affective empathy, C = cognitive empathy
Part = participant, S = student, P = pharmacist
*difference within same interview type (e.g., difference between setting, empathy, or participant)
‡difference between interview types (e.g., cognitive compare to think-aloud interview)

Proposed model of SJT response processes. Based on the findings from the analysis of prevalent codes, there was evidence to support an underlying SJT response process that can be generated from salient observations in the cognitive and think-aloud interviews. In this section, a model has been proposed that builds on the foundations of Tourangeau and colleagues (2000), who describe survey response processes as including four key components: comprehension, retrieval, judgment, and response selection. This framework was combined with features previously reported to be salient in the response process (Griffin, 2014; Lievens & Motowidlo; Ployhart, 2006) in addition to new features identified through this exploratory research.

The model, provided in Figure 6, includes the four primary components connected to the features that are proposed to influence each step in this process. Features that are bolded are those that have substantial evidence from cognitive and think-aloud interviews to support their

existence in SJT response processes, whereas those that are not bolded have limited data to support their inclusion. All features that were evaluated in this exploratory analysis were included as there were references to all components at least once in the process; therefore, the significance of these relationships cannot be excluded. A larger sample size would be necessary to confirm if the minor features could be excluded in subsequent models. Within each box connected to the primary component, features are ordered in terms of their prevalence (i.e., features that are higher on the list were referenced more frequently and identified as having a notable influence on the response process).

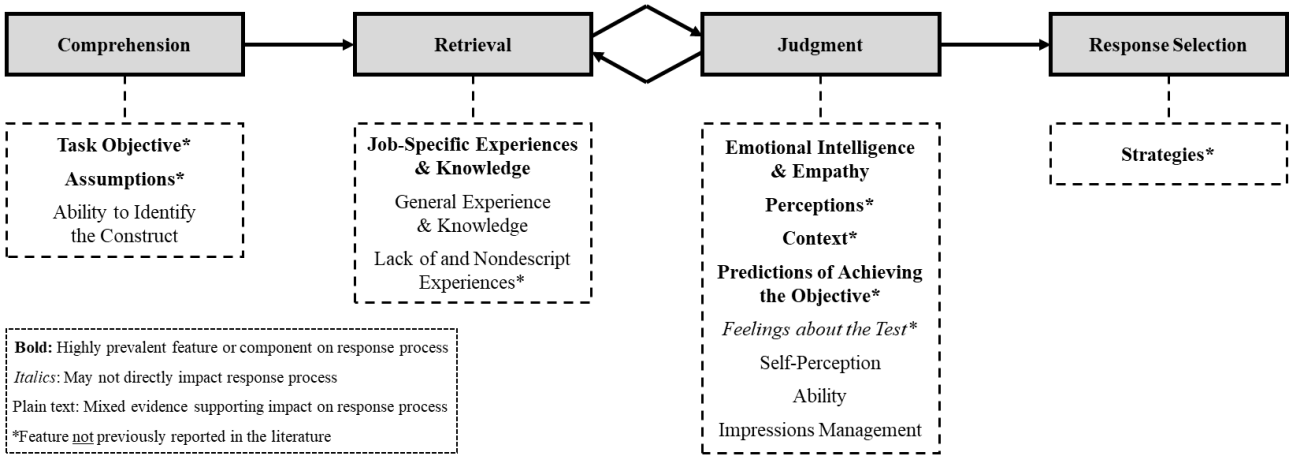


Figure 6. Proposed model of SJT response processes

In this model, the relationship between the individual components is not fully specified as the focus of this research was to explore the response process holistically. Additional research is necessary to conclude which components are most influential in SJT performance, how they relate to one another as well as other variables, and whether they influence multiple components of the response process instead of the single component structure provided here. Moreover, this model has been constructed using an SJT intended to measure empathy, therefore, this model may not be broadly applicable to other constructs evaluated using an SJT. It is, however, the first step in developing a more comprehensive and integrated model than previously documented in

the literature. The following subsections describe the pertinent features as they pertain to the four primary components of the proposed model.

Comprehension Component. Comprehension was included as a component in the model as this was considered an essential component of SJT response processes (i.e., a participant must read the item to be able to answer it accordingly). Cognitive interviews are the preferred strategy to understand comprehension in survey research (Leighton, 2017); however, questions about comprehension (e.g., difficulties with interpretation, confusion about item wording, etc.), were not included in the cognitive interview questions as that was not the focus of the research.

Comprehension also included references to how participants interpreted key elements of SJT scenarios, which is a significant component of the comprehension process. This research identified two features not previously described in the literature: (1) participants often identified a task or objective that needed to be completed and (2) participants made assumptions about the scenario. In addition, the comprehension component is connected to the ability to identify the construct as the examinee's interpretation of the item can be related to the suspected construct—this is discussed later in this chapter in relation to research question four.

Task objective identification and response prediction. An important feature observed in the comprehension process was that participants often identified an objective that was to be achieved in the scenario. Provided response options were then evaluated—in the judgment process—based on predictions of how well that response would achieve the targeted objective, among other factors. The objectives identified by participants in the cognitive and think-aloud interviews were categorized based on the goal they described. A list of these categories, descriptions, and examples are provided in Table 18; these categories are ordered from most to least prevalent.

Table 18

Categories of Comprehension Task Objectives Identified by Participants

Task Objective Category	Description	Example of Task Objective Identification	Example of Task Objective Prediction
Information Exchange	Desire to collect information or share information with another individual	"You want to finish educating thoroughly" (P07)	"You still get the information you need" (S15)
Inconclusive / General	Reference to a non-specific task or objective	"This one was a little difficult in that I didn't see an end game" (S04)	"Because that never ends well" (S15)
Emotional Improvement	Desire to positively impact feelings or avoid provoking negative feelings	"I was mostly focusing on how to help the patient best to feel better" (S10)	"This can make them more anxious" (S11)
Problem Resolution	Desire to identify or contribute to correcting an issue identified in the item	"I want to identify what can help solve this issue" (S11)	"I think if you do that well, that can really solve the problem" (S05)
Acknowledge	Desire to bring awareness to a challenge or issue	"They want you to validate their sense of loss" (P01)	"They may that you're just throwing whatever they've said under the rug" (P08)
Relationship Modification	Desire to change the interaction between two individuals	"Let them know that they can trust you" (P03)	"That would not establish rapport" (S15)

The task objective most often referenced by participants was related to the exchange of information, which could include collecting or sharing information with another individual. The objective least often described by participants referred to modifying a relationship, often between a patient and the healthcare provider. Of note, many task objectives were broad and lacked a specific focus. For example, participants made general statements about something working well or not without any indication of an explicit goal.

In general, participants discussed task objectives more often for questions related to healthcare settings compared to a non-healthcare setting and students more frequently identified objectives than pharmacists; however, none of these differences were statistically significant when compared across interview types, item classification, and participant type using Pearson's X^2-test, as reported in Table 19. This suggests that participants attempt to identify the task objective during this SJT regardless of the item setting, empathy component being assessed, or participant type.

Table 19

Frequency of References to Comprehension Task Objectives Based on Interview Type Organized by Item Classification and Participant Type

Item Classifications and Participants	Interview Type		Pearson X^2-Test	
	Cognitive (n =198)	*Think-Aloud (n = 303)*	X^2	*p-value*
Setting				
Healthcare	126	185	.34	.56
Non-healthcare	72	118		
Empathy Component				
Affective	92	146	.14	.71
Cognitive	106	157		
Participant				
Student	129	207	.54	.46
Pharmacist	69	96		

Comprehension assumptions. In addition to identifying the objective of SJT scenarios, comprehension of SJT items also included the participant making key assumptions about the presented case. Throughout the cognitive and think-aloud interviews, participants made statements about how they interpreted information that was provided. These assumptions could be classified according to what the assumptions were about, which is summarized in Table 20 with descriptions and examples. The assumption categories are organized from most to least prevalent across all interviews.

The reference to assumptions was also evaluated based on the interview type, item classification, and participant type to determine if there were patterns when assumptions may be more prevalent, as shown in Table 21. There were no statistically significant differences in the number of references to assumptions based on interview type, item setting, empathy component assessed, or participant type. There is some evidence to suggest that the type of component being assessed may contribute to varying uses of assumptions, however, the extent of this finding was not well-supported at this time and may emerge with larger sample sizes.

Table 20

Categories of Comprehension Assumptions Made by Participants During Comprehension

Assumption Categories	Description	Example of Assumptions
Person	Assumption about the actors within the scenario	"Maybe they are lying but I don't start with that – I'm not going to assume that" (S04)
Tone	Assumption about how individuals are communicating in the scenario	"It sounded really cold, just you're required to finish" (S15)
Severity	Assumption about the potential consequences or stakes associated with an outcome of a scenario or response	"Chance are if they got in front of you, it wouldn't make you late" (S01)
Information Accuracy	Assumption about if the information provided was truthful and complete	"So, if it really was an error… I would first apologize" (P02)
Urgency	Assumption about how quickly the situation needs to be addressed	"I'm going to assume it's urgent based on that I would apologize" (S04)
Position	Assumption about the relative position of the individual in the scenario	"I'm assuming in the last scenario you're not on the safety committee" (S04)

Table 21

Frequency of References to Comprehension Assumptions Based on Interview Type Organized by Item Classification and Participant Type

Item Classification and Participant Type	Interview Type		Pearson X^2-Test	
	Cognitive (n = 96)	*Think-Aloud (n = 93)*	X^2	*p-value*
Setting				
Healthcare	46	52	1.21	.27
Non-healthcare	50	41		
Empathy Component				
Affective	36	48	3.81	0.051
Cognitive	60	45		
Participant				
Student	59	59	.08	.78
Pharmacist	37	34		

In general, assumptions appeared to serve as a component of the response process for some participants when there were insufficient details provided in the scenario. As many of these scenarios were designed to exclude extraneous details, it was possible that this required more inferences by the participants. One participant, S04, best described this process as "there's a fair amount of projection" onto the scenario, depending on the elements that were provided. These data suggest that details about the scenario may be necessary if the use of assumptions in the comprehension process is not desirable. Overall, assumptions made up a small proportion of the

total number of codes (3.1%), therefore, there is minimal evidence to suggest that assumptions are an overwhelmingly significant component of the response process. It is evident, however, that assumptions can be used by participants to fill in the gaps and it may be advisable that SJT design includes explicit statements for examinees pertaining to assumptions about the setting or other features to avoid misinterpretation.

Retrieval Component. Retrieval was the next component of the response process in which participants reflected on knowledge and experiences pertinent to the scenario while they formulated their response selection. In this research, all codes referring to retrieval were also mapped onto codes that referred to job-specific and general knowledge and experiences. The significance of this component is described in greater detail in the section pertaining to RQ2. Of note, in the proposed model, there is a bidirectional relationship between retrieval and judgment that differs somewhat from the original model presented by Tourangeau and colleagues (2000). The proposed model suggests that the response process is not always linear and can integrate various memories and judgments that build on each other prior to the final decision in the response selection, which was evident by participants who retrieved multiple experiences or knowledge elements when discussing SJT items.

Judgments Component. Judgments represented the most prominent code in both the cognitive and think-aloud interviews. This included comments about the decision-making process as well as any value statement made while assessing the response options. The analysis for this component was focused on factors of SJT frameworks that pertained to the judgments, such as emotional intelligence, self-perception, ability, and impressions management. In addition, three new sub-codes were identified during the analysis: perceptions, feelings about the test, and context. Perceptions will be discussed in this section and feelings about the test will be

described in the subsequent section after the model has been described. Contextual factors that were identified are described extensively as it pertains to RQ3.

Judgments of emotional intelligence and empathy. One of the most prominent judgments included the use emotional intelligence, which was defined as the capacity to be aware of, control, and express one's emotions as well as the emotions of others. The frequencies of these references are provided in Table 22, which is organized by interview type, item classification, and participant type to determine if any patterns of use were present. Of note, the only statistically significant difference ($X^2 = 4.42$, $p = .04$) was with respect to the empathy component being assessed. According to the data, emotional intelligence was referenced more frequently for items that measured affective empathy compared to cognitive empathy in both cognitive and think-aloud interviews. This suggests that items intended to measure affective empathy may be eliciting emotional intelligence more often than those targeting cognitive empathy.

As this SJT was intended to measure participant empathy, further analysis regarding emotional intelligence focused exclusively on the participant references to affective and cognitive empathy throughout the cognitive and think-aloud interviews. Explicit references to affective and cognitive empathy, however, were relatively infrequent across interviews compared to other codes. Cognitive empathy, for example, represented 1.2% of all codes and affective empathy represented 0.9% of all codes. Table 23 includes a summary of the references to affective and cognitive empathy as it pertains to item classification and participant types studied.

There was a statistically significant difference in the presence of references according to the empathy component being assessed ($X^2 = 21.04$, $p < .001$). References to affective empathy, for example, were more common for questions assessing affective empathy whereas cognitive

empathy was discussed more often for questions assessing cognitive empathy. This finding provides additional validity evidence to support that the administered SJT items targeted specific subcomponents of empathy as intended. The data also suggested a potential difference based on the setting of the question in that affective empathy may be discussed more often in non-healthcare settings compared to cognitive empathy.

Table 22

Frequency of References to Judgment Emotional Intelligence Based on Interview Type Organized by Item Classification and Participant Type

Item Classification and Participant Type	Interview Type		Pearson X^2-Test	
	Cognitive (n = 361)	*Think-Aloud (n = 144)*	X^2	*p-value*
Setting				
Healthcare	184	62	2.58	.11
Non-healthcare	177	82		
Empathy Component				
Affective	191	91	4.42	**.04**
Cognitive	170	53		
Participant				
Student	181	83	2.32	.13
Pharmacist	180	61		

Table 23

Frequency of References to Judgment Affective and Cognitive Empathy Organized by Item Classification and Participant Type

Item Classification and Participant Type	Empathy Component Referenced		Pearson X^2-Test	
	Affective (n = 58)	*Cognitive (n = 72)*	X^2	*p-value*
Setting				
Healthcare	20	37	3.73	.053
Non-healthcare	38	35		
Empathy Component				
Affective	42	23	21.04	< **.001**
Cognitive	16	49		
Participant				
Student	26	36	.35	.56
Pharmacist	32	36		

This finding suggests that components of empathy may be more readily identifiable based on the setting of the question or the actors in the scenario; however, this cannot be confirmed. An analysis to determine if there were differences based on interview type was not included due to

the low frequency of empathy references in the think-aloud interviews. In summary, there is evidence to suggest that features of emotional intelligence were present in SJT response processes; specifically, empathy was the focus of this research, so this feature was reviewed in greater detail and showed there were some differences in the distribution of these codes consistent with the component of empathy being assessed.

Judgments of self-perception, ability, and impressions management. The remaining factors in SJT frameworks—self-perception, impressions management, and ability—were infrequently discussed among cognitive and think-aloud interviews but were still included in the model as they pertained to judgments in SJT response processes and were consistent with theoretical frameworks about SJTs. Of these three codes, self-perception was the most common, which represented 2.9% of all codes. Impression management and ability were lower, representing 1.0% and 0.3% of all codes, respectively.

Self-perceptions shared by participants often focused on either: (1) attributes of their personality (53.0% of references), (2) their identity as a healthcare provider, friend, or family member (38.8% of references), or (3) their comfort with a presented scenario (8.2% of references). References to their participant personality often included comments such as, “I think I’m probably a little bit less aggressive” as shared by P11 or S11 who discussed that, “I’m not very confrontational”. References to participant identity typically related to their status as a healthcare provider, such as P07 who stated, “I guess being a pharmacist though, it’s a little clearer”. These references also included their identities outside of work as well, such as when P03 shared that “as a new parent” there are differences in how they perceived some situations. Lastly, some participants were aware of their comfort with engaging in certain scenario; for example, S02 stated “I’d feel more comfortable talking about the error if it was something like

food". Each of these types of self-perceptions contributed to their judgements about the scenario and could impact their response selection; however, overall there was limited evidence to suggest their criticality in the process.

Moreover, there was even less evidence to support the role of impression management and ability in SJT response processes. With regards to impression management, individuals were instructed at the start of the study that the test could be used for selection into health professions programs or residency programs. When asked during the cognitive interview if that influenced their responses, an overwhelming majority of all participants (80%) noted they had forgotten about that element of the test. For the participants who did not forget, they described a struggle with differentiating their answer choices on what they should do compared to what they would do as expected by the individual administering the test. For example, S12 shared they "kind of knew what the right answer was versus what I would actually do was harder to separate". Additional research in a true high-stakes setting is warranted to further describe impressions management as it relates to SJT response processes in health professions education.

With regard to ability, participants most often made references to a lack of a knowledge of skill set that would allow them to operate best in the given scenario instead of affirmations about their abilities to succeed in a situation. For example, P07 recognized that "as a pharmacist, I'm not really trained to walk-through the risks and benefits in that case". Overall, the few references to abilities limited the analysis; however, the factor was still retained within the model as there was some evidence to suggest ability (or the lack thereof) may be play a role in the response process in that some response options were ranked lower if the participant did not feel they had the skill set necessary to successfully carry out a response option.

Judgment perceptions. Another new feature identified by the research was that participants made references to perceptions of factors weighed when evaluating response options. These perceptions were coded throughout the cognitive and think-aloud interviews, then categorized based on the features that were most salient. Table 24 includes a summary of the most prevalent categories, as well as a description and example for reference.

Table 24

Perceptions that Influenced Participant Judgments

Perception Categories	Description	Examples of Perceptions
Image	Perceptions about how the response would reflect on their image as a person	"It just makes you seem lazy" (S03)
Would / would not do	Perceptions about what the examinee would or would not do in real life	"I knew exactly what I would do there" (P02)
They want	Perceptions about what the actor in the scenario would want	"That's not what they want to hear" (P04)
Integrity	Perceptions about the honesty or legality of a response option	"You're not portraying the situation how it actually happened" (S10)
Instinct	Perceptions about what inherently feels wrong or right in the scenario	"I feel what felt right" (S02)
I want	Perceptions about what the examinee would want if they were the actor in the scenario	"I ranked these in the order that I would want somebody to do for me" (P06)

The most prevalent comment from participants was regarding the impact on their image that would follow if a certain response option was selected. Participants most frequently identified negative attributes about the impact on their image including thoughts that it could: "make you look like a jerk" (S10), "come off like accusing the patient" (S03), and "seem unprofessional" (P06). In general, there was a significant concern about how nice a response was or perceptions about the tone in which something was delivered, which could subsequently impact their image and response selection. Examples included comments about response options that "sounded really cold" (S15) or that "can come off a little harsh" (P05); these responses were then not as highly ranked. Similar to this was the perceived integrity of certain response options; for example, participants evaluated if the response was an honest reflection of the situation or if the response was legal. Each could potential have implications for the image, but these focused

specifically on an important element other than how professional or how nice they were coming across. Other perceptions included an awareness of what individuals would do in real-life scenarios, as well as a balance between perceptions of what participants believed individuals would want in the scenario along with what they would want in the scenario. Lastly, some individuals referenced their instincts in the scenarios and stated, "it just feels right" (S13) as their reasoning.

The distribution of perceptions across interviews, item classification, and participant type was also explored and reported in Table 25. There was no evidence to suggest a statistically significant difference in the frequency of perceptions based on setting, empathy component assessed, or participant type as it relates to the interview type. Overall, there was evidence that perceptions are a significant feature in SJT response processes, however, the distribution did not differ by item classification or participant type.

Table 25

Frequency of References to Judgment Perceptions Based on Interview Type Organized by Item Classification and Participant Type

Item Classification and Participant Type	Interview Type		Pearson X^2-Test	
	Cognitive (n = 164)	*Think-Aloud (n = 185)*	X^2	*p-value*
Setting				
Healthcare	87	88	.94	.33
Non-healthcare	77	96		
Empathy Component				
Affective	90	96	.25	.61
Cognitive	74	88		
Participant				
Student	108	116	.30	.59
Pharmacist	56	68		

Response selection. The last component of SJT response processes is the response selection, which included ranking response options in the format of the SJT used in this study. In general, response selection was an important element in the cognitive and think-aloud interviews as it represented 8.0% of all the codes. Response selection in this study included any reference to

the final ranking assigned to any response option. A notable feature of the response selection in this research study was the integration of general strategies that participants reported using throughout this SJT.

Response selection strategies. During the cognitive and think-aloud interviews, broadly applicable strategies used by participants in the response selection process became apparent. Table 26 summarizes the different strategies that were used by participants in making their final selections. In general, most participants approached the response process in the way they were instructed to, which was to rank responses from most to least appropriate. Others, however, considered working backwards in some situations or identifying the extremes (most and least appropriate) first and then filling in the remaining ranks. Other strategies included comparing response options, guessing, and using a process of elimination. Some participants when reading questions aloud also rephrased the item by orienting themselves within the question. One pharmacist, for example, started each response option with "Do you…" when reading the item aloud despite this not being present in the written document.

The distribution of reported strategies across interview types, item classification, and participant type was also evaluated and reported in Table 27. Overall, there were no differences in the frequency of strategies used based on the setting or the empathy component being assessed. There was, however, a statistically significant difference ($X^2 = 5.01$, $p = .03$) in the frequency of strategies used by students and pharmacists. During the think-aloud interviews, for example, students made references to strategies more often than pharmacists. In summary, there was some evidence to suggest that general test taking strategies are a relevant feature in SJT response processes and the use of strategies may differ based on who is taking the test. Additional research is warranted to determine if certain strategies are related to performance on

the test. Also, it is unclear if the distribution of strategies differs based on the type of construct being assessed or the format of the response selection in the studied SJT (e.g. ranking the response options compared to rating each response option individually).

Table 26

Strategies Used During Participant Response Selection

Strategy Categories	Description	Example of Strategies
Ordered Approach		
Best to Worst	Identify responses in order from most to least appropriate	"Going from what would be least conflict inducing to most inducing" (P11)
Worst to Best	Identify responses in order from least to most appropriate	"I started with the least appropriate and worked my way to most" (P04)
Extremes First	Identify responses at the extremes first (least and most appropriate) then the middle	"I identified the first and fifth one" (P06)
Chronologically	Identify responses in order that actions would be taken	"I would do every single one of these in this order" (P10)
Pattern	Identify responses in a type of pattern that is fairly consistent	"I'm noticing a pattern – acknowledge, ask, offer, tell, stay" (S06)
Compare Responses	Evaluate response ranking by comparing two at a time	"So, deciding between imagining things and confronting the person" (S12)
Rephrase	State the responses in a different way to identify the ranking	"So, what do I do?" (S09)
Guess	Randomly assign rankings to a response	"I just kind of put numbers down because I didn't know" (S12)
Before Reading Responses	Attempt to identify the best response before reading the options	"Before even looking at the answers, I would think about…" (S02)
Process of Elimination	Assign a ranking based on what remains after ranking others	"I guess through process of elimination it leaves…" (P07)

Table 27

Frequency of References to Response Selection Strategies Based on Interview Type Organized by Item Classification and Participant Type

Item Classification and Participant Type	Interview Type		Pearson X^2-Test	
	Cognitive (n = 68)	*Think-Aloud (n = 110)*	X^2	*p-value*
Setting				
Healthcare	28	54	1.06	.30
Non-healthcare	40	56		
Empathy Component				
Affective	42	65	.13	.72
Cognitive	26	45		
Participant				
Student	31	69	5.01	**.03**
Pharmacist	37	41		

Participant reflections about SJT processes. Throughout the cognitive and think-aloud interview, participants often shared feelings about this SJT and the response process. These feelings were reviewed and categorized into three groups: (1) effort (e.g., what made the test easier or more difficult), (2) appeal (e.g., what the participants liked and disliked about the test), and (3) thoughts about the response options. Table 28 provides a summary of these features, a description, and example from the transcripts.

Table 28

Features of Participants Reflections about SJT Processes

Reflection Features	Description	Example of Participant Reflections
Effort	The ease or difficulty of the item or test	"I thought this one was hard" (P01)
Appeal	Test elements that were liked or disliked	"I sort of hate answers like this" (S04)
Response Options		
Similarities	Comments about how similar response options were to one another	"The answers were a little bit similar" (S07)
Outlandish	Comments about how ridiculous or preposterous a response option was	"So that seems like an odd answer now that I read it" (S06)
Desire to Combine	Comments about wishing to include to responses together instead of rank	"I wish I could have combined or wish I could have tied" (S03)
Missing	Comments about a response option that was desired to be included but wasn't	"I wish there was an option on here that said…" (S14)
No Right	Comments that there were no right answers in the options provided	"I don't think there's a right answer" (P06)

The most prominent feeling about this SJT was that the questions were more difficult than expected; 99 comments were made by participants about the difficulty compared to 61 comments about the ease of answering the questions. Feelings about the test may not be highly relevant in SJT response processes, however, they can be important elements of validity and design research about SJTs. In this case, participants identified features that made the examination harder or easier and what contributed to their perceived success on this SJT. There was limited consensus, however, on which features made the test easier or difficult. Most often, participants referenced prior experiences as a salient factor that made items easier. One pharmacist (P06), for example, shared that, "ones that related to more a personal experience I think were easier to answer, where I've been in that situation and could better answer based on

what went well or didn't go well". Another pharmacist, P11, simply noted that the mere requirement of "ranking them one through five was hard". Forcing participants rank response options may elicit different feelings about the test compared to others; therefore, additional research should be considered as to how design elements affect feelings about the test if this is a pertinent concern. Overall, the data suggest that participants frequently struggled with the examination due to the complexity of the design and the task to be completed.

Summary of RQ1. In summary, there is evidence that SJT response processes include four key components: comprehension, retrieval, judgment, and response selection. This is the first research that has explicitly shown these components are present and to offer a model that integrates these features from the literature with evidence. Moreover, this research identified the factors that contribute to each of these components in the response process. Five new features not included in previous research were identified in this study: identification of the task objective, assumptions about SJT scenarios, perceptions of the response options, contextual features of the response options, and general strategies in the response selection. In general, these features were consistent across the item characteristics and participant types tested. The results from this research question greatly expand on our current understanding of SJT response processes and offer a model to frame future research to generate validity evidence for SJTs.

RQ2: Role of Experience in SJT Response Processes

The second research question was: "*What is the role of job-specific experiences and knowledge in the response process to SJT items?"* Specifically, the goal was to explore if participants referenced different types of experiences and knowledge that was pertinent to their process when ranking the answer choices. Lievens and Motowidlo (2016) have suggested that SJTs integrate job-specific as well as general knowledge and experiences, but the extent to which

these features are integrated in SJT responses has not been studied. In addition, this question explored the retrieval component of the response process model presented in the previous section. Overall, this question was aimed at determining whether certain job-specific or general knowledge or experiences are retrieved more often in SJT response processes.

This section: (1) summarizes the data analysis process for the second research question, (2) reviews the prevalence of codes related to experience and knowledge, and (3) describes salient features of experiences and knowledge shared by participants.

Data analysis summary. During the cognitive interviews for each SJT item, 10 students and 10 pharmacists were asked if they thought of any experiences when answering the test question and to describe those experiences. Participants also may have referenced experiences and knowledge at other points in the cognitive and think-aloud interviews without prompting. Transcripts were coded to identify job-specific knowledge and experiences as well as general knowledge and experience as defined in the codebook (Appendix L). The codebook was modified during the initial review of transcript data by the researchers to include two additional codes, which included references to *nondescript experiences* (i.e., the researchers could not clearly identify if the experience was explicitly connected to a healthcare setting or not) and references to a *lack of experience* (i.e., participants not being aware of knowledge or experiences related to the scenario).

Coded segments were aggregated and quantified across items and participants to investigate if there was a pattern regarding the reference to job-specific or general knowledge and experiences in SJT response processes. Participants had to make at least once reference to either of these knowledge or experiences during the cognitive interview to be counted in the data analysis.

Prevalence and distribution of codes related to experience and knowledge. In summary, of the 480 participant references to knowledge and experiences throughout the cognitive and think-aloud interviews: 45.2% related to job-specific knowledge or experience, 27.5% related to general knowledge or experience, 17.9% related to a lack of experience, and 9.4% were nondescript experiences. The SJT for this research study, however, included an equal number of items pertaining to healthcare and non-healthcare settings; therefore, it was anticipated that both types of knowledge and experiences would be described equally. In this case, the unequal distribution of job-specific and general experiences suggests that participants either use varying degrees of job-specific and general knowledge and experiences when responding to SJT items or that participants may integrate job-specific knowledge even in non-clinical scenarios. Conversely, the distribution of comments was sufficiently equal between students and pharmacists with 51.0% of the references by students and 49.0% of the references from pharmacists, which suggests that individuals may recall information and experiences pertaining to the same fields regardless of their level of experience. Overall, these findings warranted further exploration to determine when job-specific knowledge and experiences were considered more applicable for SJT items and to evaluate if there were differences in the type of knowledge and experiences recalled by participants based on their level of experience (i.e., students and pharmacists).

Prevalence according to item characteristics and participant type. First, participant responses for the cognitive interviews were aggregated to determine the number of participants who reported which type of knowledge or experience was relevant to SJT test items. These data, provided in Table 29, were compiled according to: the setting of the question (i.e., healthcare or non-healthcare), the empathy component being assessed (i.e., affective or cognitive empathy),

and the participant type (i.e., student or pharmacist); this was to determine if one type of knowledge or experience was more prevalent based on any of these factors.

Pearson's X^2-test was conducted to determine if there were significant differences in the frequency that job-specific knowledge and experiences or general knowledge and experiences were recounted in relation to the previously described classifications. There was a statistically significant difference in the reference to job-specific and general knowledge or experiences based on whether the item was in a healthcare or non-healthcare setting ($X^2 = 73.62$, $p = < .001$); in this case, job-specific knowledge or experiences were referenced more often than general knowledge and experiences when the setting was healthcare related whereas general knowledge and experiences were more commonly cited when items referred to a non-healthcare setting.

There was also statistically significant difference in the reference to job-specific and general knowledge or experiences based on whether the item measured affective or cognitive empathy ($X^2 = 14.52$, $p = < .001$); the data suggest that job-specific knowledge and experiences are referenced more frequently by participants when answering questions intended to measure cognitive empathy compared to those intended to measure affective empathy. These results suggest that the construct being assessed can have implications on the type of experiences and knowledge recalled. Conversely, there was no statistical difference in the number of participants who identified job-specific and general knowledge and experiences reported by students compared to pharmacists ($X^2 = 1.63$, $p = .20$); this suggests that the participant type did not relate to differences in the overarching classification of experiences they recalled, which further confirms the initial observation.

Table 29

Frequency of Participants who Reported Job-Specific or General Experiences and Knowledge during Cognitive Interviews Organized by Item Characteristics and Participant Type

Item Classification and Participant Type	Job-Specific		General		Total	
	Experience (n = 106)	*Knowledge (n = 86)*	*Experience (n = 114)*	*Knowledge (n = 9)*	*Job-Specific (n = 192)*	*General (n = 123)*
Setting						
Healthcare	87	56	27	4	143‡	31‡
Non-healthcare	19	30	87	5	49‡	92‡
Empathy Component						
Affective	49	34	60	6	83‡	66‡
Cognitive	57	52	54	3	109‡	57‡
Participant						
Student	50	42	62	6	92	68
Pharmacist	56	44	52	3	100	55

Notes: ‡ $p < .001$, all other comparisons statistically non-significant ($p > .05$)

Prevalence according to SJT item. To better observe the relationship between the types of knowledge and experiences recalled during this SJT, the frequency that participants described job-specific and general knowledge or experiences was further classified at the item level that is summarized in Tables 30a and 30b. Fisher's exact test was used to evaluate if there were any differences in the frequency which students or pharmacists referred to job-specific and general knowledge and experiences during each SJT item. There were no statistically significant differences between the two groups for all but two of the items (*p*-values ranged from .06 to 1.00), which is consistent with the previous findings.

Item CN3 had a statistically significant difference ($p = .03$) in the number of pharmacists who referenced job-specific experiences when asked how they should respond to a scenario where they believe the patient is lying about their diabetes management; in this item, all 10 pharmacists referenced job-specific experiences compared to half of the students. Item AN2 also had a statistically significant difference ($p = .02$) between pharmacists and students in that students were more often reported not having any experience when working with individuals who were having difficulty conceiving a child. Each of these examples show that, although

infrequent, it is possible that some SJT items can be designed to target experiences that may not be encountered equally among examinees. The implications of this finding, however, are limited.

Table 30a

Frequency of Participants who Reported Job-Specific or Nondescript Experiences and Knowledge in Cognitive Interviews Organized by SJT Item

Item	Job Experience		Job Knowledge		Nondescript Experience	
	Pharmacist (N = 15)	*Student (N = 15)*	*Pharmacist (N = 15)*	*Student (N = 15)*	*Pharmacist (N = 15)*	*Student (N = 15)*
CH1	8	8	7	8	2	2
CH2	6	8	2	2	5	2
CH3	9	9	7	6	0	1
CN1	5	3	2	4	1	1
CN2	7	9	9	4	4	2
CN3	10*	5*	3	2	0	1
AH1	2	1	7	7	5	1
AH2	1	1	1	2	1	2
AH3	4	0	2	1	0	1
AN1	4	2	2	2	1	1
AN2	0	2	0	0	1	2
AN3	0	2	2	4	0	0
Total	56	50	44	42	20	16

Note: *p < .05

Table 30b

Frequency of Participants who Reported General or a Lack of Experiences and Knowledge in Cognitive Interviews Organized by SJT Item

Item	General Experience		General Knowledge		Lack of Experience	
	Pharmacist (N = 15)	*Student (N = 15)*	*Pharmacist (N = 15)*	*Student (N = 15)*	*Pharmacist (N = 15)*	*Student (N = 15)*
CH1	2	0	0	0	2	5
CH2	1	1	0	0	0	3
CH3	1	3	0	1	2	5
CN1	3	6	1	0	4	3
CN2	2	3	1	1	5	3
CN3	1	4	0	0	1	2
AH1	9	9	1	1	3	4
AH2	7	7	0	0	3	4
AH3	7	7	0	0	6	4
AN1	2	6	0	1	6	5
AN2	8	6	0	2	1*	7*
AN3	9	10	0	0	3	2
TOTAL	52	62	3	6	36	47

Note: *p < .05

Salient features of experiences and knowledge. After investigating how references to experiences and knowledge were recalled across this SJT, the next step was to investigate the salient features of the experiences and knowledge that were referenced. Coded transcripts were first analyzed to determine if there were consistent features of the experiences and knowledge referenced by participants overall. References to job-specific and general experiences often included features related to the location, the actors, and the task or topic. In addition, the experiences could be classified on their similarity to the presented scenario, the specificity of the details provided, and the recency of the memory to the present moment. Features of knowledge references included information, a strategy, or a skill that was applicable to the scenario. Table 31 provides a description of these features and examples from the transcripts.

Table 31

Features of the Experiences and Knowledge Referenced by Participants during this SJT

Features of Experiences and Knowledge	Description	Example of Experiences and Knowledge
Experience		
Location	The setting of the experience	"I was called to a different ICU and the patient had an infusion that had been running at the wrong rate" (P11)
Actors	The individuals included in the experience	"I've had patients before that have complained to me" (P05)
Task / Topic	The challenge or goal of the experience	"I think anytime you have patients who are upset… you can relate it back to your own experiences" (P06)
Similarity	How consistent the memory is with the presented scenario	"I don't think I've been in a situation very similar to this" (S10)
Specificity	The level of details provided about the experience	"I remember as a resident doing something right, being told by a nephrology resident…" (P10)
Recency	The amount of time between the memory and the experience	"Just actually two days ago, the patient we had was on Harvoni…" (P07)
Knowledge		
Information	Facts or observations pertinent to the situation	"This one had me immediately thinking about the legal implications of a medication error" (P03)
Strategy	A plan or approach to achieve an objective	"I want to ask them—why they think that, why they want to do that and tell them to talk to their doctor" (S12)
Skill	An ability or set of strategies to achieve an objective	"I just thought about my training… when it comes to our service with hard motivational interviewing" (P14)

Description of job-specific experiences. With respect to job-specific experiences, pharmacists and students generally referenced these elements in very similar ways with some

notable exceptions. The most substantial difference between the two groups was the location of the referenced experiences. Most pharmacist examples of job-specific experiences were explicitly connected to their work; very few of their references included examples from pharmacy school. Conversely, job-specific experiences shared by students had a larger variety and included experiences from school, clinical rotations, and some work experiences. The greater distribution is likely due to the recency of these experiences for students compared to practicing pharmacists; therefore, pharmacists with more experiences are likely to rely on their work-based experiences more so than experiences that were from their earlier years of training.

In addition, student experiences more often included observations of interactions in which they were not an active participant as well as shared stories, class discussions, and simulations. For example, S10 discussed how they had "seen some pharmacists delivering sensitive information about what could happen with certain drugs"; a pharmacist, P13, when discussing the same test item instead thought "about a situation when [they] were practicing in the HIV clinic". Another example was from S3 who stated, "I know we talked about a lot of different scenarios in class… especially diabetes patients" and S2 who shared, "we've talked about medication errors in class a lot and I've talked about it on some of my rotations". The data suggests that students more often integrate job-specific experiences that relate to their education and training witnessed so far, which may not include their direct involvement in a similar scenario.

Moreover, when pharmacists discussed job-specific experiences, they often included a greater amount of detail about the scenario compared to students who tended to be more generic in their descriptions. P7, for example, shared a story that "two days ago, that patient we had that was on Harvoni, it was documented in the clinic notes" and continued to describe in detail the

experience of identifying a medication error. Students, on the contrary, are less descriptive with similar scenarios. For example, S6 talked about an experience that included "going into the patient's room when the patient's family is upset at something" and had difficulty recalling many details about the event. In general, the data suggests that when pharmacists do provide an example they often include additional details and elements compared to students.

Description of general experiences. The use and quality of general experiences, however, was not significantly different between pharmacists and students. In general, the experiences tended to be somewhat vague but still closely related to the presented SJT scenarios. The actors in these scenarios were often friends and family members and the discussion about these experiences occurred mostly when discussing items referring to non-healthcare settings. One notable feature was that examples from televisions shows were sometimes referenced as viable experiences. For example, when P15 was discussing the item related to a friend taking a medication to help them study their immediate response when asked about the question was "Jesse Spano – from Saved by the Bell". One student, S13, also discussed "I think of experiences that a lot of times I watch on TV shows like Dateline". Overall, there is minimal evidence to suggest that general experiences include particularly salient features that contributed to SJT response processes differently based on the level of participant experiences.

Description of job-specific knowledge. The references to job-specific knowledge were also consistent between students and pharmacists. The majority of job-specific knowledge references related to information as described in Table 31. Information often included specifics about disease state management, facts about specific medications, the legality of certain actions, and references to hierarchical structures in healthcare. An area of difference, however, was in the skills that were frequently referenced by pharmacists and students. For example, many

pharmacists referred to service recovery training they had received, which P6 described as "when you have a situation that has escalated and how it is best to handle it". In this study, there were no students who referred to training that was similar as this skill that was taught to pharmacists in the workplace. Conversely, few pharmacists referred to mental health aid training, which was more often discussed by students. S14 described how mental health first aid training "explicitly emphasized that you shouldn't talk about yourself in mental health crises and you should really be focused on addressing that person's need and affirming them". In this study, there is evidence to suggest that there are minimal differences in the types of knowledge participants use to answer SJT items regardless of their level of experience; however, there may be some nuances based on organizational requirements and shifts in classroom education over time.

Description of general knowledge. Compared to job-specific knowledge references, the discussion of general knowledge was very limited. When discussed, general knowledge often referred to information such as social norms such as "just thinking about social norms, you wouldn't confront somebody in the grocery store", as shared by S14. In summary, there were few conclusions that could be drawn regarding the use of general knowledge as it appeared infrequently in participant transcripts. The scant presence suggests general knowledge may not be a substantial component in SJT response processes.

Description of nondescript experiences and knowledge. Nondescript experiences were not analyzed as few conclusions could be made from the references. Examples included instances where P1 stated "this [question] is a tough one because I feel like this like a reality every day" and S14 who shared "this one felt familiar to me". References to a lack of experience, however, were reviewed to determine if they were more prevalent in specific scenarios. There were no differences in the number of references in healthcare and non-healthcare settings (41 to

42 references, respectively) but there were more references to a lack of experience when completing questions intended to measure affective empathy (48 statements) compared to those measuring cognitive empathy (35 statements). In addition, it was shown previously that students and pharmacists can differ in the number of participants who admit to a lack of experience.

Description of lack of experience and knowledge. Overall, there were minimal differences in how participants referred to or how they perceived their lack of experience. Most participants, like S3, stated "I don't really have very much to draw on" or simply "this has never happened" as shared by P14. One difference, however, was that pharmacists tended to be more specific when they considered whether they had experiences to draw from. For example, P6 stated "I haven't had a particular scenario with regards to chemotherapy" whereas students discussing the same question would state more generally that they "haven't been in a situation where a family member is that upset" (S3). The data suggest that pharmacists may be more attentive to granular details compared to students when searching for similar experiences.

References to experiences and knowledge in think-aloud interviews. Lastly, references to job-specific and general knowledge and experiences in the think-aloud interviews were analyzed to identify prominent patterns. Overall, there were few references to job-specific and general experiences, general knowledge, nondescript experiences, or the lack of experience during the think-aloud interviews (i.e., five or less participants making a reference to any component). Job-specific knowledge was referenced by 12 pharmacists and nine students. A majority (91%) of the references to job-specific knowledge were related to information, such as disease state management, the legality of response options and the responsibilities as a healthcare provider. The remaining references were regarding strategies for engaging with patients, such as apologizing and sharing experiences in times of emotional stress. Overall, there was scant

evidence to verify the role of knowledge and experiences in the this based on the think-aloud interviews, except that job-specific knowledge may be explicitly involved.

Summary of RQ2. In summary, there is evidence to support that job-specific and general knowledge and experiences are a significant component of SJT response processes. Of note, data from the cognitive interviews show that job-specific references are more prevalent, regardless of the setting of the test item and that there can be significant variation in the job-specific experiences retrieved by participants. Experiences often include features such as the location, actors, task, similarity, specificity, and recency whereas knowledge can be classified by information, strategies, and skills. There is minimal evidence, however, to suggest that experience and knowledge are explicitly referenced during the response process according to the think-aloud interview data. Overall, these findings contribute substantially to SJT research in that this was the first attempt at generating evidence about the types and features of experiences and knowledge that are recalled during SJT response processes.

RQ3: Role of Setting in SJT Response Processes

The third research question was: *"What is the role of the setting presented in SJT items in the response process (i.e. the influence of a healthcare or non-healthcare specific settings)?"* Specifically, the goal was to explore how participant responses would have changed based on a different setting. In addition, the aim was to describe contextual features considered by participants during the response process. Lievens and Motowidlo (2016) argue that SJTs may not be as contextually specific as previously thought and there is suspicion that the situational elements of SJTs may not be critical. The results pertaining to this question, therefore, offer evidence about the significance of the item setting and the implications for SJT design and response processes.

This section includes: (1) a summary of the data analysis process for the third research question, (2) the perceived impact of a change in item setting (e.g., switching the setting from a healthcare to non-healthcare setting), (3) a description of the contextual features believed to influence response selections, and (4) a comparison of setting features shared during think-aloud interviews.

Data analysis summary. The focus for this research question was to explore the relationship between the setting and SJT response processes. During the cognitive interview for each item, 10 students and 10 pharmacists were asked whether their ranking of the response options would have changed if the setting was switched (e.g., changed to a healthcare setting if the question was in a non-healthcare setting). Participants answered affirmatively or negatively to the question and then were asked to provide reasons for their choice. The frequencies were reported for each item and summarized based on item characteristics (i.e., setting and empathy component assessed) and participant type (e.g., students and practicing pharmacists). To describe which factors about the setting may contribute to SJT response processes, transcripts from the cognitive interviews and think-aloud interviews were screened for comments about how participant answers would change depending on specific factors. The prevalence of these features across items characteristics and participant type was also compared to identify salient patterns among cognitive and think-aloud interviews.

Perceived impact of a change in item setting. In summary, participants stated that their selected responses would change secondary to a change in the setting 51.3% of the time (123 affirmed / 240 requests); this suggests that item setting contributes to SJT response processes and requires further exploration.

Impact of setting based on item characteristic and participant type. First, participant responses were aggregated to determine the number of individuals who reported their responses would change. These results were compiled in Table 32 according to: the setting of the initial question (i.e., healthcare or non-healthcare), the empathy component being assessed (i.e., affective or cognitive empathy), and the participant type (i.e., student or pharmacist); this was to determine if a response was more likely to change in the context of item setting, empathy component assessed, or participant type.

Table 32

Frequency and Comparison if a Change in Setting Affects Response Selections by Item Characteristic and Participant Type

Item Classification and Participant Type	Setting Affects Response		Pearson X^2-Test	
	Change (n = 123)	*No Change (n = 117)*	X^2	*p-value*
Setting				
Healthcare	61	59	.02	.90
Non-healthcare	62	58		
Empathy Component				
Affective	52	68	6.02	**.02**
Cognitive	71	49		
Participant				
Student	67	53	2.02	.16
Pharmacist	56	64		

Pearson's X^2-test was conducted to determine if there were significant differences in the frequency that a change in setting was reported to prompt a change in the response selections. There was a statistically significant difference for items measuring affective and cognitive empathy ($X^2 = 6.02$, $p = .02$) in that participants were more likely to report their responses would change as a result of a shift in the setting for items that measured cognitive empathy compared to items that measured affective empathy; this suggests the response process may be influenced to a greater extent depending on the construct being measured. In this case, it can be interpreted that measures of cognitive empathy may be more susceptible to differences in the setting presented in SJT items and that cognitive empathy is not equally applicable across various settings. In other

words, understanding the perspectives of others may vary based on the setting or contextual elements provided. Conversely, there were no statistical differences in how often a response would change based on the initial setting of the item or the participant type.

Impact of setting based on SJT item. The frequency that a change in setting may influence response processes was further classified at the item level, which is summarized in Table 34. Fisher's exact test was used to evaluate if there were any differences in the frequency which students or pharmacists identified whether a change in the setting would prompt a change in their response; there were no statistically significant differences between the two groups for any of the items. Participants were also asked how their response would change; these comments were reviewed by the researcher and summarized in Table 33 to identify if there were notable differences between student and pharmacist responses (refer to Appendix C for the test items, if needed). Overall, for each item there were at least four participants who stated that a shift in the setting would lead to a change in their responses, which further supports that the setting can impact the response process for participants regardless of the item.

Pharmacists and students often reported similar approaches in how their responses would change based on a shift in the setting. In summary, there was a mixture of cases in which there were differences between pharmacist and student responses that were: substantial (e.g., CH2), subtle (e.g., CH1, CH3, AH1, AH2, CN3, and AN3), and consistent (e.g., AH3, CN1, CN2, AN1, and AN2). This distribution suggests that shifting from a healthcare to non-healthcare setting can lead to differences in how students respond to scenarios compared to pharmacists; conversely, shifting from a non-healthcare setting to a healthcare setting has more consistent changes in the response for students and pharmacists.

Table 33

Frequency of When a Change in Setting Affects Response Selection by SJT Item and How the Response Changes by Participant Type

	Identified Setting Significance*			How the Response Changes	
	Pharmacist (N = 10)	*Student (N = 10)*	*Total (N = 20)*	*Pharmacist*	*Student*
Healthcare				*If the setting were non-healthcare related, they would…*	
CH1	6	9	15	Agree with the observation Notify the person	Share personal stories more Notify the person
CH2	4	0	4	File a complaint earlier Ask for alternative sources	***
CH3	5	7	12	Confront them about lying Ask fewer questions	Confront them about lying
AH1	2	6	8	Share personal stories more	Share personal stories more Recommend a professional
AH2	3	3	6	Not explain the cause	Not explain the cause Apologize earlier
AH3	7	9	16	Transition from the topic sooner / stop talking	Transition from the topic sooner / stop talking
Non-Healthcare				*If the setting were healthcare related, they would…*	
CN1	5	4	9	Instruct them not to take the medication earlier	Instruct them not to take the medication earlier
CN2	10	10	20	Not allow the patient to cut the line	Not allow the patient to cut the line
CN3	5	6	11	Divert conversation earlier Dismiss the family sooner	Divert conversation later Leave the location
AN1	3	5	8	Request more information Not readily leave	Request more information Not readily leave
AN2	4	4	8	Discuss treatment options Not discuss experiences	Discuss treatment options Not discuss experiences
AN3	2	4	6	Support decision earlier	Support decision earlier Recommend a professional
TOTAL	56	67	123	***	***

Notes: Group differences were statistically non-significant according to Fisher's exact test ($p > .05$)

The one item with substantial differences (CH2) between responses had zero students reporting that they would change their responses, whereas four pharmacists stated they would change their response. This question referred to difficulty when gathering a medication history for a patient over the phone with another pharmacist. When asked how participants should respond to this scenario when in a non-healthcare setting, all 10 students stated they would approach the problem in the same way, however, several noted they would be more likely to give up on obtaining the information if it was for personal reasons alone. For example, one student

shared, “what I would probably do is just be like, “Oh, okay” and hang up” (S02). Some pharmacists shared that they would not necessarily handle the situation as calmly; one suggested they would be “a lot grumpier” (P01) and one stated that “if they were rude, I would probably file a complaint more often” (P05). This example suggests shifts in certain settings can have greater influence on the response process depending on the participant type, however, this was the only case in this study.

Another unique example was item CN2, in which all 20 participants reported their responses would change if the setting were switched to a healthcare context. This item, which refers to a woman asking to cut in line at the grocery store to get home to her sick child, was perceived differently when applied to a healthcare setting in which a patient was asking to cut in front of someone. All participants in this case referenced rules or policies in healthcare that prioritize patients based on the severity of the situation, which may not be as susceptible to change as seen in non-healthcare settings. Participants discussed how they “triage in the emergency department” (P03) or use “transplant waiting lists” (S02) as examples to describe how patients are screened accordingly and placed in an order that is not often modified. One pharmacist stated there is a “protocol that you can fall back on” (P06) , which made the decision much more straightforward. This example suggests that certain settings are more conducive to rules or strategies that may not be broadly applicable; therefore, further supporting that the setting can play a significant role in SJT response processes.

The remaining items included few or subtle differences in how participants would change their response based on a shift in the setting. Item CN3, for example, asks participants how they should address a situation in which a sibling is being questioned regarding their marital status. When asked if this question was switched to a healthcare setting in which a patient was being

asked numerous questions in front of them, this prompted varying responses from pharmacists and students that may be attributable to factors such as their level of experience or comfort with these cases. Pharmacists often reported that in the healthcare setting they would divert the questioning much earlier and would be willing to be dismissive of the family more so than they would have engaged in a non-healthcare setting. Specifically, one pharmacist, P06, stated they would be "making sure that the family members understood that the patient is the important priority" and that they would be willing to step in if the patient is visibly uncomfortable based on their authority as a healthcare provider. Conversely, some students decided to change their responses because they were more apprehensive about intervening with patient's families; for example, student S11 suggested "I don't want to step into their argument because that's their life". This example demonstrates that a change in the response may not be consistent among participant types; therefore, understanding how the response would change can illustrate how other factors about the setting can contribute to SJT response processes.

Description and distribution of the setting features. To describe which factors about the setting may contribute to SJT response processes, transcripts from the cognitive interviews and think-aloud interviews were screened for comments about how participant answers would change depending on specific factors. Interestingly, there were 175 uses of the phrase "it depends" (and other equivalents) by participants across the transcripts, which signified the importance of contextual elements in SJT response processes. These factors were coded and classified into four categories: (1) factors pertaining to the participant or *examinee*, (2) factors pertaining to *actors* in the presented scenario, (3) factors pertaining to the *relationship* between the examinee and actors, and (4) factors pertaining to the *situation*. Table 34 outlines the factors

grouped by the four categories and more specific examples of these categories are provided in Table 35.

Participants often cited multiple factors that influence their response process and that these factors could affect their response differently based on the scenario. For example, item AH1 asks how the participant should respond to a patient who is upset about the recent loss of a loved one. One pharmacist, P06, stated that "If it were a friend, I would have been more inclined to share my own personal experiences…I'd feel more comfortable sharing personal loss and talking about it on a more personal level". In this case, the participant identified that the actor (e.g., a friend instead of a patient) has an impact on the response selection as well as the relationship (e.g., a personal instead of a professional relationship).

Table 34

Factors about the Item Setting Perceived to Influence SJT Responses Grouped by Category

Classification of Setting Features	Example Setting Factors that Influence Response
Examinee	Position and responsibilities of the examinee (e.g., healthcare provider, manager)
	Needs, wants, or expectations of the examinee (e.g., responsibilities for patient care)
	How the examinee portrays emotion or communicates to an actor (e.g., how something is said)
	How long the examinee would pursue a specific action or outcome (e.g., interest in the goal)
	Proximity of the examinee to the situation (e.g., directly involved / affected, observing)
Actor	Position and responsibilities of an actor (e.g., healthcare provider, family, friend)
	Needs, wants, or expectations of an actor (e.g., what the patient would want)
	How an actor portrays emotion or communicates to the examinee (e.g., how rude they are)
	How an actor responds or is anticipated to respond to an action (e.g., potential outcome)
	An actor's personality (e.g., openness, willingness)
Relationship	Relative position between examinee and actor (e.g., boss, student, sibling)
	Whether the relationship is expected to be professional OR personal
	How long an actor and examinee have known one another (e.g., duration of the relationship)
	How much information an actor and examinee know about one another (e.g., likes, history)
	Level of comfort between examinee and actor (e.g., comfort with being honest)
	How information is shared between examinee and actor (e.g., medium of communication)
Situation	Severity of the consequences related to an action (e.g., high-stakes, low-stakes)
	Severity of the current situation (e.g., safety, emotional wellbeing, necessity)
	Legal or liability of potential actions or lack thereof (e.g., false documentation, illegal drugs)
	Actions that were previously attempted (e.g., other steps taken prior to the question)
	Amount of information available or that could be obtained (e.g., background knowledge)
	Feasibility or capability to complete potential actions (e.g., authority, resources)

Table 35

Examples of the Setting that Affect Responses by SJT Item

Item	Pharmacist Examples of Setting Influences	Student Examples of Setting Influences
CH1	*Relationship:* "because you have more of a relationship with that person" (P08)	*Relationship*: "I'm a little more reticent to share a personal story as a healthcare provider" (S04)
CH2	*Actor*: "If the pharmacist…is really abrupt and abrasive then it changes how you respond" (P04)	*Situation*: "When it has to do with medications… then it gets to be a little higher stake" (S07)
CH3	*Relationship*: "I think depending on my relation to that person, I would act accordingly" (P09)	*Relationship*: "With a family member you just already have that trust" (S13)
AH1	*Relationship*: "I know this individual personally" (P03) *Actor*: "It's not as straightforward… depending on where the patient is in their disease state" (P04)	*Relationship*: "It's going to skew your decision… what is your connection with them" (S04) *Actor*: "it depends on the patient… the patient is usually thinking about this more" (S01)
AH2	*Situation*: "I can fix that. I can say "You don't want the hamburger, okay" (P01)	*Situation*: "I'd feel more comfortable talking about the error if it was something like food" (S02) *Examinee*: "Am I in a position of responsibility in this setting?" (S04)
AH3	*Situation*: "In a healthcare setting where there are policies and procedures to follow" (P06)	*Situation*: "It also depends on the hospital policy" (S13) *Relationship*: "It would depend on… how close of a friend it was" (S02)
CN1	*Examinee*: "That can come off a little harsh so that's why… to make it softer" (P05)	*Situation*: "You could buy caffeine pills… but Adderall is a controlled substance" (S12)
CN2	*Situation*: "Is this a new job presentation versus is this just your study thing" (P04)	*Situation*: "Standing in line at a checkout is a lot lower stakes than a healthcare setting" (S07)
CN3	*Examinee*: "The part of us that's a little bit gossipy or curious will say outside our context" (P14) *Relationship*: "Depending upon the level of relationship with that patient" (P15)	*Relationship*: "We're not all bunch of friends… like I have some sheen of authority" (S02) *Situation*: "You don't know all the facts… so that vagueness would make it difficult" (S06)
AN1	*Actor*: "It depends on which parent… because I like one better than the other" (P02)	*Situation*: "I guess I just assumed that there was some kind of worst case scenario" (S06)
AN2	*Examinee*: "I was the bitch with the baby" (P01)	*Examinee*: "If I were their healthcare provider? Then I'll be just really clear" (S11)
AN3	*Examinee*: "It would be different depending on your views of higher education" (P06)	*Examinee*: "You're thinking about how involved you want to be in this person's situation" (S03)

Participants commonly identified that relationships with friends and family members come with different expectations compared to relationships with work colleagues or patients. For example, student S10 shared that when trying to convince a patient about not taking a non-prescribed medication compared to convincing a friend, they thought "it'd be easier because you could come at it from the standpoint of I've had training in this… and there's no evidence to back this up or that's illegal". In this case, factors such as the examinee's training as well as the legality of the situation also contribute to the response process. Altering the question to exclude factors such as the illegality or the examinee's position could alter their response.

Pertinent setting features based on SJT item. The distribution of these factors across the 12 SJT items was also investigated to determine if there were potential patterns related to the component of empathy being assessed, the setting of the item, or the participant. Participant references to the factors were aggregated and the most prevalent categories of factors are provided in Table 36 with supplementary examples for reference. In general, there was no discernible pattern related to which categories were more prevalent based on each SJT item. Student and pharmacists often agreed on the salient factors that could influence their response; however, students were more likely to list multiple factors compared to pharmacists for each item.

References to setting features in think-aloud interviews. Lastly, think-aloud interviews were reviewed to determine if participants thought that their responses would change based on elements of the setting as previously described. In general, there were few references by pharmacists and students to how their response selection would depend on these factors. Of the four categories, participants most often cited factors related to the situation, such as the necessity for more information, the severity of the scenario, and the severity of consequences related to a specific action. It was expected that the think-aloud interviews would provide minimal data as participants were not explicitly prompted to consider changes in the setting during the exam process. Moreover, imagining alternatives to the presented scenarios would be considered unproductive in answering the questions. Data from the think-aloud interview therefore, provides little evidence to support how the setting contributes to SJT response processes except that participants occasionally remarked that their responses may change due to factors presented in the scenario.

Summary of RQ3. In summary, there is evidence to suggest that the setting and contextual features of an SJT item may have a role in the response process. Data from the cognitive interviews show that participants identified their response to an item would have changed if it were in a different setting more than half of the time they were asked. Features that affected their responses were classified into four groups including factors related to the: examinee, actors in an SJT item, relationships between the examinee and actors, as well as additional elements about the situation. In addition, there was no discernable pattern that identified when certain factors would be more salient. Lastly, there was minimal evidence that participants actively consider how the setting contributes to alternative response options during the examination process according to the think-aloud interview data. Overall, the results of this research question provide evidence that the setting of SJT items may affect the response process according to participant beliefs, suggesting that SJTs are likely not exclusively tests of general domain knowledge and skills.

RQ4: Role of the Ability to Identify the Construct in SJT Response Processes

The final research question in this study about SJT responses processes was to describe the role of the participant's ability to identify the construct being evaluated. The final research question was: *"What is the role of the ability to identify the construct being evaluated (i.e. empathy) in the response process to SJT items?"*. Specifically, the goal was to explore what participants believed each item was measuring and how that related to their performance on this SJT. Griffin (2014) presented the first argument about the significance of the ability to identify the construct as it pertains to examinee performance on other instruments used in the health professions. The relationship of this ability to SJT performance, however, has not been

evaluated. The results of this research question were intended to determine if the ability to identify the construct has a significant role in the response process.

This section: (1) summarizes the data analysis process for the fourth research question, (2) identifies the frequency of participants who identified the construct of interest (i.e., empathy), (3) outlines other constructs identified by participants, and (4) describes challenges participants shared about identifying the constructs.

Data analysis summary. During the cognitive interviews for each question, 10 students and 10 pharmacists were explicitly asked to describe what knowledge or ability the item was measuring. Responses were reviewed by the primary researcher and categorized based on the construct identified. For the purposes of this research, participants were required to explicitly state "empathy" as the construct being evaluated to be categorized into that group (i.e., no synonyms for empathy were permitted); this approach minimized the potential for misinterpretation by the researcher and served as a conservative estimate for exploratory purposes. The other constructs were reviewed, however, to determine if they may be appropriate to include into this estimate. Moreover, if a participant did not identify empathy as the construct being evaluated during the cognitive interview, they were asked to further describe their answer to help classify their response during the analysis phase. Cognitive and think-aloud interviews were also coded when participants made attempts to identify the construct of interest outside of being explicitly asked.

Frequency that empathy was the identified construct. Participants specifically identified "empathy" as the construct being assessed 33.3% of the time (80 out of 240 total responses across 12 items), which was the construct most frequently identified across the entire

test. Of note, compassion was also a frequently identified construct that could be considered synonymous with empathy; therefore, empathy or synonymous construct was identified 35.6%.

Frequency that empathy was identified by item characteristic and participant type.

Table 36 provides a summary of how often empathy was reported as the construct being measured group by setting (i.e., healthcare or non-healthcare), empathy component (i.e., affective or cognitive), and participant (i.e., student or pharmacist). For these calculations, only the cases identified exclusively as empathy (i.e., the 33.3%) were included as the addition of compassion was not a substantive increase in the initial finding. In total, pharmacists identified empathy as the construct being assessed less often (27.5% of the time) than students (39.2% of the time). When empathy was identified as the construct being measured it was most often reported for items in a non-healthcare setting (56.3%) rather than a healthcare setting (43.7%) and for questions targeting affective empathy (71.3%) rather than cognitive empathy (28.7%).

Table 36

Frequency and Comparison of Participants Identifying Empathy as the Construct Being Assessed by Item Characteristic and Participant Type and the Correlation to Score on the Item

Item Classification and Participant Type	Identified Construct		Pearson X^2-Test		Empathy Identified-Item Score Correlation	
	Empathy (n = 80)	*Not Empathy (n = 160)*	X^2	*p-value*	r_s	*p-value*
Setting						
Healthcare	35	85	1.87	.17	.07	.44
Non-healthcare	45	75			.12	.20
Empathy Component						
Affective	57	63	21.68	**< .001**	.11	.23
Cognitive	23	97			.03	.72
Participant						
Student	47	73	3.68	.78	.03	.06
Pharmacist	33	87			.12	.19

Notes: Spearman correlation (r_s) used to examine relationship between whether the participated identified empathy as the construct and participant score on the item

Pearson's X^2-test was conducted to determine if there were significant differences in the frequency empathy was identified by item characteristic and participant type. There was a

statistically significant difference in the frequency empathy was reported for items measuring affective or cognitive empathy ($X^2 = 21.68$, $p < .001$) and no difference based on the item setting or the participant type. Spearman's correlation coefficient was also calculated to explore the relationship between identifying empathy as the construct being assessed and participant scores on the respective item based on item characteristic and participant type; there were no statistically significant relationships.

Frequency that empathy was identified based on SJT item. The frequency that empathy was identified as the construct being measured can be further classified at the item level, which is summarized in Table 37. In addition, Spearman's correlation coefficient was calculated to describe the relationship between identifying empathy as the construct and the participant score on the respective item. Fisher's exact test (not reported in the Table) was used to evaluate if there were any differences in the frequency which students or pharmacists identified items as measuring empathy; there were no statistically significant differences between the two groups.

Table 37

Frequency that Participants Identified Empathy as the Construct Being Assessed by SJT Item and the Correlation to Item Score

	Pharmacists Identified (N = 10)			Students (N = 10)			All Participants (N = 20)		
Item	*Empathy*	r_s	*p-value*	*Empathy*	r_s	*p-value*	*Empathy*	r_s	*p-value*
CH1	2	.63	.05	4	.29	.42	6	.39	.09
CH2	1	–.44	.20	1	.00	.99	2	–.12	.62
CH3	1	.22	.55	0	***	***	1	.02	.93
CN1	1	.12	.74	1	–.31	.38	2	–.08	.75
CN2	3	–.04	.91	3	.35	.32	6	.17	.49
CN3	3	.00	.99	3	–.15	.67	6	–.06	.81
AH1	8	–.09	.80	9	.24	.50	17	.04	.87
AH2	0	***	***	3	–.21	.57	3	–.06	.79
AH3	2	–.09	.80	4	.00	.99	6	–.10	.66
AN1	1	–.18	.14	4	.11	.76	5	.09	.70
AN2	8	.14	.70	9	.12	.74	17	.10	.67
AN3	3	.00	.99	6	.37	.29	9	.17	.47
TOTAL	33	–.13	.64	47	.28	.32	80	.11	.56

Items reported to measure empathy most often were AH1 and AN2, with 85% of participants identifying empathy as the construct being assessed. These items, however, were not correlated with QCAE scores as reported previously in Table 12 (r_s = –.01 and –.14, respectively). The three items that were least often reported to measure empathy were CH3, CH2, and CN1, with only 5%, 10%, and 10% of participants identifying empathy as the construct being measured, respectively. Correlations of item score and identifying the construct were not statistically significant for any SJT item, which suggests that the ability to identify the construct may not be related to SJT performance assuming the test was adequately measuring empathy.

Other constructs identified by participants. SJT items were reported to measure a variety of other constructs including conflict management, integrity, and teamwork, which were identified 15%, 12%, and 9.6% of the time. Table 38 provides a summary of which constructs were most prevalent based on each SJT item. For the items least often identified to measure empathy (CH3, CH2, and CN1), the constructs reported by participants varied. For CH3, the item was most likely measuring gathering information and conflict management, whereas CH2 was reported to measure teamwork. Conversely, 75% (15/20) of participants identified CN1 to be measuring integrity, which suggests this item was not measuring the intended construct.

Challenges with identifying constructs. At the end of the cognitive interview participants were asked whether they believed their responses would have changed if they had known initially that the entire test was intended to measure empathy. Most participants (10 students and 11 pharmacists) stated their answers would not or would probably not change; one student S05 confirmed "that [it] was pretty easy to see in the questions" and a pharmacist, P03, stated that they "picked up on that anyway". One participant (P09) even shared that "To be

honest though, I wanted to say empathy for all of them, but I felt like I couldn't [laugh]", which suggests some participants may have provided other guesses about what they thought the item measured due to their assumption that the test had to be measuring multiple constructs. The remaining nine participants suspected several of their answers may have changed, which one student S07 shared that "it would have been very easy to just look for the most empathetic answer" if they were aware prior to the test.

Table 38

Participant Reported Constructs Measured Summarized by SJT Item

Construct Reported	*CH1*	*CH2*	*CH3*	*CN1*	*CN2*	*CN3*	*AH1*	*AH2*	*AH3*	*AN1*	*AN2*	*AN3*	*Total*
Conflict management	1*	2*	5	3	2	10	2*	2‡	4	3*		2*	36
Integrity	2*			15	4			1‡	4	2		1*	29
Teamwork	3	6	2*		2			8				2	23
Compassion	4	3	2*			3	1‡			1*	2	5	21
Adaptability	1‡	2‡			1‡			1*	2*	6	1*	1*	15
Prioritization	1‡	1*			5			1‡	2				10
Professionalism	1*	1‡	2			1*		1*	2‡	1‡			9
Gathering information			7										7
Critical thinking		2	1‡					1*		2‡			6
Customer service	1*	1*						2*					4
TOTAL	14	18	19	18	14	14	3	17	14	15	3	11	160

Note: *Construct only reported by pharmacists, ‡Construct only reported by students

Of note, participants often struggled during the cognitive interview when asked what the item was measuring. Several student and pharmacist participants shared their frustration stating, "I don't like this question" (S15) and that "it's really hard, deep" (P09). Often, they summarized the task to be completed in the item instead of defining what the question was measuring, which required further probing. In addition, multiple participants requested for "a list of knowledge and abilities" (P11) that could help them in the process, which was not provided. Difficulty in addressing the question about identifying the construct being measured suggests that individuals may not often consider this factor when taking examinations such as an SJT.

This was further supported by a lack of evidence during the think-aloud interviews that suggested participants actively concentrate on what the item is intending to measure. Of the 30 participants, only two participants (both pharmacists) speculated about what the question was targeting during the examination. Specifically, one pharmacist (P02) mentioned "this is something then I guess about professionalism, how do you empathize with them?". The other pharmacist (P03), was more generic in their remarks and tended to summarize the task such as "this is a scenario where you need to communicate with another professional" and "this is a scenario where your job is to relay information". The three utterances were not sufficient to suggest that identifying the criteria being assessed is at the forefront of thought during the examination.

Summary of RQ4. In summary, there was minimal evidence to support that the participant's ability to identify the construct being measured has an appreciable relationship to performance on an SJT or SJT response processes. Empathy was the construct that was most often identified when participants were asked what an SJT item measured; however, most of the constructs reported by participants were not explicitly connected to the targeted construct of interest. In addition, participants may be able to identify the construct more readily based on the subcomponent being assessed (i.e., affective or cognitive empathy). In general, participants struggled when identifying the construct being measured, which suggests their attempts to identify the construct may be inaccurate and that this process is not commonly conducted by examinees. Lastly, there was no evidence to suggest that participants actively attempt to identify the construct being measured during the examination process based on the think-aloud interview data. Overall, these results describe the first attempts of how the ability to identify the construct may relate to SJT performance.

Summary

Results of the quantitative and qualitative analyses provide evidence that SJT response processes include the complex integration of comprehension, retrieval, judgments, and response selections, which has not been comprehensively explored in the literature. In addition, the results identified salient features of the response process and identified new features not previously described in SJT research. There was evidence to suggest that job-specific experiences and knowledge comprise a significant portion of the retrieval process and that SJTs target job-specific elements as suspected. Moreover, there was evidence that supports the notion that SJTs are highly contextual and that changes in the item setting or factors can impact response selections. Lastly, there was inconclusive evidence of how the ability to identify the construct being assessed relates to examinee performance. Overall, the findings make significant contributions to the understanding of SJT response process and offer substantial implications for the validity evidence of interpreting SJT scores.

Chapter 5: Discussion

This chapter begins with a discussion of the significance and implications of the results presented in Chapter 4 as it pertains to each of the research questions as well as the overall study. The chapter continues with a discussion of the challenges with measuring professional competence, challenges with research on response processes, and the challenges with designing and conducting research on SJTs. The chapter concludes with a consideration of the limitations of the present study and proposed directions for future research.

Significance and Implications of Results

Discussion of the psychometric properties of this SJT. The presented research intended to address gaps in the literature regarding the response process in SJTs. To address these questions, it was essential to design an SJT that targeted a single construct—failure to create such an instrument could introduce confounding factors that would limit the interpretation of the findings. For example, one of the research questions focused on how well examinees could identify the construct being tested. If examinees identified a construct different than the one intended, it would be interpreted that examinees were incorrect; however, the finding may be the result of a poorly designed instrument that did not measure the desired construct, therefore, their responses would not be inaccurate. The evidence-centered design approach, also referred to as a construct-driven SJT by Lievens (2017), was used to create the instrument because it offered a systematic approach to define the construct and ensured alignment between the test items and construct of interest (Mislevy, Almond, & Lukas, 2003; Riconscente, Mislevy, & Corrigan,

2016). Analysis of the psychometric properties of the administered SJT was a foundational step in the research to provide evidence that this SJT was measuring the intended construct and, therefore, support subsequent interpretations of the results.

The compiled SJT score data were negatively skewed with substantially positive kurtosis, which indicated that participants performed well on this SJT. In general, there was variation in SJT total scores, however, variation in participant performance can only be used to describe relative standing on the construct being measured. In other words, higher SJT scores were simply indicative of greater participant empathy compared to other participants and there were no comparisons to a normative sample or population. The limitation of this is discussed in subsequent sections about challenges measuring professional competence and areas of future research. Ideally, it would be desirable to identify if there are certain thresholds that could be indicative of a "sufficient amount" of empathy that would be linked to effective patient-provider relationships or positive patient outcomes.

On the one hand, the low observed Cronbach's alpha values were concerning as this suggested that a unidimensional construct was not being assessed. Indeed, an internal consistency index such as alpha may be inappropriately applied when the data result from a measure that is intentionally multidimensional. Alternatively—and perhaps more likely—the low values may be attributable to the small numbers of items, highly variable inter-item correlations, and homogeneity of the participant sample. On the other hand, low Cronbach alpha values are not uncommon in SJT research. For example, Catano and colleagues (2012) conducted a meta-analysis of 39 studies and identified a weighted mean corrected reliability coefficient of .46. In this light, the low alphas found in this study are consistent with existing research about the multidimensional nature of SJTs (Lievens, 2017). For these reasons, the low alpha levels were

tentatively considered to be acceptable and did not prohibit additional analyses. Conversely, the positive correlation of SJT scores with the QCAE (r_s = .34, p = .07) provided additional evidence to support that this SJT was measuring empathy as intended. Overall, the results of the psychometric analyses suggested that the SJT developed for this study provided a reasonable estimate of participants' empathy given the constraints of a small sample size and brief scale (i.e., 12 total items).

Discussion of RQ1. The first research question ("*What factors and strategies are involved in the cognitive processes when examinees respond to SJT items?"*) was posed to investigate the key features of SJT response processes, which until this study had been significantly under-researched (Krumm et al., 2015; Rockstuhl et al., 2014). To address this question, think-aloud interviews and cognitive interviews were used to elicit the response process during and after the administration of an SJT. Of note, the think-aloud interviews were particularly important to this research question as they were conducted prior to the participants being asked about their thought process during the exam; in other words, the think-aloud interview was most likely to be reflective of the natural response process that was unadulterated by the questions posed by the researcher. Emphasis, therefore, was placed significantly on the think-aloud interview findings as it pertained to this research question.

Prior research about survey response processes suggested that the cognitive process during an SJT would be similar; this was suspected to include elements of comprehension, retrieval, judgment, and response selection (Ployhart, 2006; Tourangeau, Rips, & Rasinksi, 2000). Results from RQ1 provided evidence that these four components are indeed present in SJT response processes according to utterances in the cognitive and think-aloud interviews; therefore, this four-component structure served as the foundation for the proposed model of SJT

response processes. More specifically, statements related to judgements, retrieval of memories or information, and response selections were some of the most prevalent codes in the qualitative analysis of the data. Overall, the findings suggest that the four-component model is an appropriate and well-supported approach to describing SJT response processes. In addition, these features can independently contribute to the decision-making process and can therefore influence score interpretations if any of these features are inappropriately influenced.

Research on SJTs had also identified multiple antecedents suspected to influence response selections. Lievens and Motowidlo (2016) shared a framework that included features such as job-specific and general experiences as well as knowledge that contribute to response selection. In addition, this framework included other individual characteristics (e.g., emotional intelligence, ability, personality, etc.) that were expected to influence the decision-making process. Results from RQ1 suggest there are a host of factors that are considered by participants during the response process, which can vary greatly among examinees in the extent to which they are applied. Specifically, results from RQ1 confirmed that job-specific experiences and knowledge as well as emotional intelligence were salient features of SJT response processes. Other features such as general experience and knowledge, self-perceptions, ability, and impressions management were not sufficiently supported as pertinent components of the response process as previously expected. The proposed model still included all of these features as the lack of utterances about a particular feature was not considered to be sufficient evidence to discard it in this exploratory phase; instead, additional research is necessary to confirm the findings of this research.

Results pertaining to RQ1 also identified new features of the response process that have not been previously described in the literature. Specifically, results from RQ1 suggested that

participants often attempt to identify a task objective during this SJT and evaluate how well response options achieve that task based on their comprehension of the elements that are presented. In addition, they often make assumptions about the scenario that influence how they comprehend the situation. During the judgment process, participants identified that they evaluated response options according to their perceptions on how the action would reflect on their image, whether it was something they could imagine doing in real life, or what they would want done for them in the situation. Moreover, participants identified that contextual features such as the item setting could greatly influence their response selections. Lastly, there were a host of test-taking strategies that participants employed during this SJT that may be broadly applicable regardless of the item.

In general, these new features have not been extensively discussed in prior research about SJT response processes. Rockstuhl and colleagues (2014), for example, were the first to report evidence about SJT response processes; however, they categorized participant utterances simply on the content presented. For example, they identified that most comments during SJT responses were about the intentions, emotions, or thoughts as it pertained to the presented scenario. This research extends on this prior work in addressing how these features relate to the four-component model of the response process that was evident and describing these features in greater detail. In addition, Krumm and colleagues (2015) presented a small research study that identified some of the strategies test-takers used when completing an SJT. Similar to their findings, the results from this study showed that strategies such as comparing response options were often cited by participants during the process. This study took that research further by identifying additional strategies and better describing how participants evaluated the effectiveness of response options (for example, by comparing how well the response option achieved the task objective the

examinee had identified). The previous work presented by Rockstuhl (2014) and Krumm (2015) was limited in the depth of information it provided about SJT response processes; the results of this research question, therefore, greatly expanded the overall understanding of these features within the response process.

Lastly, analyses were conducted to determine if these features of the response process and model differed substantially based on item characteristics (e.g., setting or empathy component assessed) and participant type (e.g., students or pharmacists). Results from RQ1 suggest there are only slight differences in the response process that may occur as it pertains to these variables. This was the first research study that examined how these components, especially differences in experience levels of examinees, may influence the response process. Therefore, the findings suggest that a general SJT response process model may be applicable; however, this research only include participants from the field of pharmacy in one region and it used a test intended to measure only one construct. The approach and model may be used though to investigate if the model is applicable to other health professions and constructs tested using SJTs in future research.

Overall, results from RQ1 were the first that explicitly showed which components of the response process were salient using cognitive and think-aloud interviews. In addition, the results were used to generate a model that can be tested through future research and be used as a mechanism to generate validity evidence for SJTs.

Discussion of RQ2. The second research question (”*What is the role of job-specific experiences and knowledge in the response process to SJT items?”*) was intended to elaborate on the retrieval component of SJT response processes by describing what types of knowledge and experiences were recalled during this SJT. The framework presented by Lievens and Motowidlo

(2016) suggested that SJTs require examinees to integrate job-specific as well as general knowledge and experiences, but the extent to which these factors were incorporated in the response process had not been studied. In addition, it was unclear if there was variation in the extent to which these factors are incorporated based on differences in the item characteristics or participant type.

The results from RQ2 demonstrated that job-specific knowledge and experiences were more often referenced by participants than general knowledge and experiences during both the cognitive and think-aloud interviews. Overall, there was scant evidence that general knowledge contributed substantially to the response process; therefore, it could be argued that general knowledge may not be a significant feature to include in future studies about the response process. This distribution of knowledge and experiences, however, may differ when testing other health professions or different constructs—additional research is needed to confirm this finding.

Of note, the test incorporated an equal number of items that were in healthcare and non-healthcare settings in an effort to elicit knowledge and experiences that were not exclusively related to a healthcare job. The higher prevalence of references to job-specific knowledge and experience suggests that participants may use varying degrees of job-specific and general knowledge and experiences when responding to SJT items. Conversely, it is possible that participants may simply integrate job-specific knowledge even in non-clinical scenarios. For example, it's plausible that much of the individual's identity is connected to their work as a clinician and those experiences are more readily accessible due to the substantial amount of time spent in those settings; as a result, participants may readily integrate those features into their collective decision-making processes. The study did not include questions or analyses that could

explicitly identify why this observation was present and it warrants further investigation in the future.

The results from RQ2 also contributed two new findings: participant awareness of a lack of experience or knowledge and descriptions of the types of knowledge and experiences retrieved. The findings demonstrated that participants often identified times in which they had little to no experience or knowledge about a particular topic, which was unexpected. The awareness of a lack of knowledge or experience was, therefore, included in the proposed response process model as a notable feature. The study did not investigate explicitly how this awareness contributed to SJT performance—it is unclear if this is a significant feature that should be considered in validity studies or when interpreting SJT scores. The findings also contributed substantially to the literature as this was the first study to describe qualities of the experiences and knowledge that were referenced by participants during this SJT. The elements, such as the location, actors, and tasks, should be considered during SJT design processes as they were identified by participants to be relevant components of SJT scenarios. These features may also be particularly relevant if there is a certain type of knowledge or skill that is to be evaluated by the test item.

Lastly, the results from RQ2 evaluated if there were differences in the types of knowledge and experiences that were recalled based on individual items, their characteristics, and participant type. As expected, job-specific knowledge and experiences were more often referenced in healthcare setting questions compared to non-healthcare setting questions, which suggested these items were capable of targeting job-specific knowledge and experiences. Interestingly, there was no evidence to suggest that students and pharmacists differed in the prevalence of job-specific knowledge and experiences that were recalled; however, there was

evidence to suggest that pharmacists more often recalled work-based experiences while students recalled classroom-based or learning experiences. In addition, there were few cases where there were differences in the knowledge or experiences retrieved for an individual item. This finding suggests that response process validity data are not necessarily generalizable and that the response processes being elicited can be sample dependent. Currently, there is a lack of research to corroborate these findings as often the research is focused exclusively on the sample of interest. Overall, this suggests that response process validity data should be interpreted with caution and be discussed in reference to the sample being evaluated.

In summary, findings from RQ2 were the first confirm the types of knowledge and experiences that are most often retrieved during SJT response processes. Job-specific knowledge and experiences remain the most salient features retrieved, which was expected considering SJTs are grounded in human resources research and were designed as an instrument to predict job performance (Campion, Ployhart, & MacKenzie Jr., 2014; Chan & Schmitt, 2002).

Discussion of RQ3. Similar to RQ2, the third research question (*"What is the role of the setting presented in SJT items in the response process (i.e. the influence of a healthcare or non-healthcare specific setting)?"*) investigated a specific feature of SJT response processes—the setting of the item. Specifically, the results were intended to address how situational SJTs are as this has been a debated topic in the literature. Krumm and colleagues (2015), for example, evaluated the stems of SJTs and determined that descriptions were not necessary for most of the items as participants would respond similarity even when scant details were provided in the scenario. In addition, Lievens and Motowidlo (2016) argue that SJTs may not be as contextually specific as previously thought and that SJTs may be tapping more general domain knowledge rather than knowledge that is specific to a scenario.

Results from RQ3 contradict Krumm's (2015) findings and the opinions of Lievens and Motowidlo (2016)—the results suggest that the setting and contextual features of an SJT item have significant impacts on SJT response processes. In the study, participants reported that a change in setting would lead to a change in their response selections over 50% of the time; therefore, the setting and contextual features of an SJT item should strongly be considered as a salient design feature that could influence participant responses.

Additional research is warranted as there are several possible explanations for this observation that differs from previous research. The first is that there is truly an impact of the setting or context of the question that may not have been accurately identified in previous studies. The second potential explanation is that participants may suspect that their response may change but may not actually be aware of the influence of the setting until they are prompted about it, which is further supported by the fact that few participants mentioned elements about the setting during the think-aloud interviews that did not include prompting questions. The third potential explanation is that previous research on SJTs has focused largely in human-services fields outside of healthcare (e.g., selection of managers, retail employees, military personnel, etc.)—it is possible the setting may be more influential in a healthcare setting due to the stakes of the consequences (i.e., life or death). Participants often discussed the balance of personal and professional interactions that are expected with healthcare providers, which may not be applicable to other settings where SJTs have been used. Moreover, non-healthcare related SJTs may be more likely to include questions that involve interactions with strangers, which is not frequently the case in healthcare related SJTs, which can be another contributing factor. In this study, very few questions included interactions with individuals who were (hypothetically) complete strangers to the participant.

In addition, there was evidence to suggest that setting may be more influential depending on the construct being assessed. In this study, the number of participants stating that their responses would change if the setting was altered differed when the item assess affective versus cognitive empathy. This finding suggests that subtle differences in the construct being assessed may influence how the setting is significant in the response process. Additional research should, therefore, investigate how the setting influences responses when evaluating other constructs and in different professions to determine if this finding is generalizable to all SJTs.

Results from RQ3 also contributed significantly to the literature as participants were requested to describe how their response would change as a result of the setting change to articulate what salient features should be considered during SJT design. Participants described key elements of the question included factors pertaining to themselves as examinees, the actors in the scenario, the relationship between those individuals, and factors related to the scenario. The findings suggest that details about these features of the test item can be influential in crafting their response to an SJT item and that the weight of each of these factors may vary substantially based on the item or the individual completing the test. These factors may contribute to construct irrelevant variance and, therefore, may influence score interpretation if they are not considered during the design process.

Overall, findings from RQ3 are the first to challenge the idea that setting and contextual features of SJT items do not significantly influence the response process. Evidence from this study suggests that details pertaining to the scenarios presented can be critical in SJTs used in healthcare settings. In addition, this research question was the first to identify the salient features of the setting that may influence the response process and should be the focus for future research evaluating the impact of changing these elements in SJT items.

Discussion of RQ4. Lastly, the fourth research question (*"What is the role of the ability to identify the construct being evaluated (i.e. empathy) in the response process to SJT items?"*) was included in the study to investigate a new area of interest in selection research—the ability to identify the construct being assessed (also referred to as ATIC). Griffin (2014) was the first to describe that the examinees who are able to correctly identify the construct being evaluated were more likely to score higher on MMI prompts in medical school admissions. This capability, however, has not been evaluated with respect to other selection strategies such as an SJT. The results of RQ4 were the first to explore how the ATIC may be related to SJT performance and to describe what participants thought items were measuring.

In this study, participants only identified empathy as the construct being assessed 33% of the time, which is likely attributed to several factors. The most critical explanation is that several of the items were likely not measuring empathy—which is consistent with the low Cronbach's alphas previously reported. If the items were measuring a unitary construct, the alphas would have been larger, and the identified construct would should have been more consistent across items. In addition, it was clear that some items were perceived to measure completely different constructs than anticipated—for example, 75% of participants stated that one item was measuring integrity instead of empathy.

Moreover, participants struggled significantly when asked to identify what the item was evaluating, with several requesting for a list of constructs to pick from. Social desirability was another factor that could explain the observed result; several participants noted that they wanted to say that empathy was the construct being assessed by multiple questions, however, they felt they had to say something different as they did not think it was plausible that the test would be measuring one construct exclusively. Lastly, participants were required to explicitly state

"empathy" as the construct being assessed to be considered correct; this was done to ensure there was no ambiguity in whether they correctly identified the construct—allowing the use of synonyms (such as compassion) had minimal effect on the overall number of people who identified it appropriately.

The variability in the constructs identified by participants also suggested that empathy may be a difficult construct to assess independent of other constructs. Across all items, conflict management was frequently cited by participants as the knowledge or skill being assessed by the question. This finding gave rise to the idea that empathy may be a construct that is often present in the setting of moments of conflict, such as those surrounding integrity, teamwork, compassion, and adaptability (i.e., the other most frequently identified constructs). If empathy requires these other elements to be present in a scenario, it limits the ability to create and interpret a unidimensional measure of empathy as there will invariably be other confounding constructs being measured. The relationship between ATIC and SJT performance must, therefore, be interpreted with caution.

A difference was also observed in that items pertaining to affective empathy were more likely to be identified as measuring empathy compared to items that were designed to measure cognitive empathy. The observation—although not unsurprising—is likely attributable to the nature of affective empathy, which evokes emotions in individuals and can be readily associated with empathy in these settings. This finding is important in that it suggests certain constructs may be more readily identifiable than others and this can have implications for interpreting the significance of the ATIC as it relates to performance. There may be less significant of a relationship to performance if an overwhelming majority of participants is able to readily identify it compared to a construct that is not as recognizable. When considering the role of

ATIC in the response process, researchers should therefore consider overall how easy of a construct it is to identify. Moreover, this research involved the use of a measure where only one construct was measured. The findings presented here related to ATIC may differ when multiple constructs are measured in an instrument and, therefore, there are multiple possible constructs that may be identified; further, the presence of multiple constructs may result in interactions that may make straightforward interpretations related to ATIC more difficult.

In this study, there was a weak positive relationship between ATIC and SJT performance. This would suggest that ATIC may not be a salient feature in the response process; however, it is likely the test was not sufficiently accurate in measuring empathy. In other words, if this test was not measuring empathy as intended, or solely, then we are unable to determine if an examinee correctly identified the construct being measured. In addition, previous work on the ATIC was conducted in high-stakes testing environments, which were not present here. Although participants were told to imagine that this test were being used for student or resident selection in a health professions program, many noted they forgot about that element and this could have affected the results in that participants may not have actively attempted to identify the construct. The results presented here related to ATIC, therefore, may vary if the use or consequences of performance on an SJT differed.

In summary, results of RQ4 provide minimal evidence that ATIC is a substantial feature of SJT responses or that it is related to SJT performance. There were several confounding factors that likely contributed to this finding, therefore, the results of RQ4 should be interpreted cautiously. Additional research is needed to evaluate the role of ATIC in the response process and the relationship to SJT performance to confirm if these findings were accurate.

Concluding remarks and implications of the results. Results of the four research questions exemplified that SJT response processes are a complex process consistent with the four-component model (i.e., comprehension, retrieval, judgment, and response selection) that has been used to describe survey responses. The results provided evidence that address two prominent questions in SJT research. First, it was determined that job-specific experiences and knowledge were more prevalent than general experiences and knowledge, which suggests that SJTs are not simply measuring general domain knowledge as previously suspected. In addition, the studied showed that the setting of an SJT item can have significant implications in the response process, which challenges previous research that suggested setting had a less prominent role.

Although not the focus of this study, conclusions about the overall validity of the SJT scores yielded by the measure developed for and used in this study are possible based on a synthesis of several sources of evidence. Validity evidence based on test content was obtained through subject matter experts who aided in developing and assessing the questions that were grounded in our theoretical understanding of empathy in the context of the health professions. Moreover, participants were asked to identify the construct during the examination, which demonstrated there were variable perceptions about what the test was measuring. In addition, validity evidence regarding the relationship to other variables (e.g., the QCAE) demonstrated a positive—albeit small—correlation in the scores suggesting a similar construct was being assessed. Cronbach's alpha was calculated to provide validity evidence about the internal structure of the SJT, which suggested the instrument was not exclusively unidimensional.

The central focus of this research was *not* on the overall validity of scores for the SJT developed here; rather the focus was on the collection and evaluation of evidence based on

response process—a source of validity evidence not routinely gathered or evaluated in previous SJT research. The research conducted here supports the notion that SJTs require complex decision-making processes. Of note, this interpretation is limited to the confines of the sample and SJT studied—additional research is needed to determine if these findings are consistent with SJTs measuring other constructs (e.g., adaptability, integrity) that are administered in other professions (i.e., healthcare and non-healthcare related).

The results have at least one significant implication: extreme caution should be considered when applying SJTs to health professions education. The variable validity evidence supporting score interpretation also limits the potential use of these instruments without additional research to corroborate whether SJTs can produce sufficiently valid and reliable results in high-stakes learning environments. For example, SJTs are used currently for student and resident selection and being considered as strategies to evaluate training and monitor progress of clinicians throughout their development; however, SJTs should not be relied upon as the only instrument to assess professional competency in these settings without further proof that they contribute valid and reliable information.

Challenges with Measuring Professional Competence and Empathy

This study sought to explore SJTs as an assessment strategy to measure a critical component of professional competence—empathy. There were significant challenges when measuring empathy that were encountered during this study that are shared within this section.

Difficult to define professional competence using strictly unidimensional constructs. In general, assessing professional competence is a formidable challenge because there are variable conceptualizations in the literature (Epstein & Hundert, 2002; Goldstein et al., 2006; Li, Ding, Zhang, Lie, & Wen, 2017). Moreover, each of these conceptualizations include multiple

subcomponents to define professional competence, which results in highly interconnected relationships between the constructs that comprise the domain. The framework used for this study, for example, defined professional competence according to nine components that outlined the knowledge, skills, and abilities necessary to function optimally as a healthcare provider beyond clinical competence (Patterson, Ashworth, Kerrin, & O'Neill, 2013). The large number of components makes assessment of professional competence a significant undertaking and requires that researcher untangle the relationships to distinguish the constructs from one another.

This challenge is further exacerbated in that many of the constructs that comprise the overall domain are also poorly defined and often share similar features. This study, for example, focused on empathy as the construct of interest. The definition of empathy varies substantially in the medical education literature and constantly shifts based on emerging research in the field and other disciplines (Hojat, 2007; Quince, Thiemann, Benson, & Hyde, 2016; Tamayo, Rizkalla, & Henderson, 2015). Moreover, the definitions of non-cognitive constructs such as empathy often differ only slightly from other constructs such as compassion, sympathy, or emotional intelligence and empathy integrates broad skills sets related to communication, problem-solving, and critical thinking (Hojat, 2007; Quince, Thiemann, Benson, & Hyde, 2016); overall, this makes distinguishing the singular construct difficult. This overlap can introduce construct-irrelevant variance that can be difficult to minimize or account for—in this study, an evidence-centered design was an approach used specifically to reduce this potential while creating a construct-driven SJT that attempted to measure one construct exclusively (Lane, Raymond, Haladyna, & Downing, 2016; Lievens, 2017). It must be noted, however, that instruments measuring the same construct related to professional competence could vary substantially based on the definition of the construct that was used.

It can also be easy to assume that each of the constructs comprising professional competence would be relatively equal in terms of their difficulty to assess; however, the experiences during this research suggest otherwise. Empathy, for example, is particularly challenging because of the overlap with other skill sets. Other components, such as integrity or adaptability, may present greater or lesser assessment challenges to the extent they are grounded in greater or less explicit decision-making processes or have more or less complex theoretical underpinnings. By extension, the application and design of SJTs to measure these constructs, may be more or less influenced by contextual features, assumptions, or other features that were identified to be significant in this study.

This challenge was also evident in this study based on the psychometric analysis and comments from participants. The findings, for example, included low Cronbach alpha values that suggest a unidimensional construct was not being assessed. Moreover, participants did not identify that empathy was being assessed for most of the questions—instead, they perceived SJT questions measured a myriad of other constructs such as conflict management, teamwork, and adaptability. Researchers, therefore, must be cognizant that designing instruments to measure professional competence require clear definitions of the constructs to minimize overlap with similar components of the domain. In addition, measures of internal consistency (such as Cronbach's alpha) may not be ideal to evaluate the reliability of SJTs. Researchers should consider other strategies such as factor or dimensionality analyses with larger sample sizes or test-retest reliability to determine the stability of SJT scores over time.

Poor understanding of construct gradients and interpreting results. Another challenge in this area of research is the difficulty in interpreting the results of measures of professional competence without a greater understanding of how movement along the spectrum

of a construct relates to desirable outcomes, performance, or behaviors. In other words, as it pertains to this study, this challenge relates to understanding what it means practically to have "a little empathy" compared to "a lot of empathy".

In the health professions, for example, assessment strategies such as SJTs are used to measure components of professional competence as a screening tool for admissions. The notion is that individuals with higher standings (i.e., higher scores) on pertinent constructs are expected to perform better in school or be more effective clinicians (Bardes, Best, Kremer, & Dienstag, 2009; Patterson, Cleland, & Cousans, 2017). The limitation is that the correlation of these variables only accounts for how these instruments rank individuals against one another; it does not necessarily provide criterion-referenced or diagnostic information about the individual. For instance, it is unknown if cut-off scores could be generated to delineate groups of students that may be at risk of poor performance in school or that may be more likely to be changed with negligence as a practitioner. Conversely, it is possible that very high scores related to empathy may have negative consequences in the event the individual is more susceptible to burnout or unnecessary stressors. Overall, there were challenges in understanding how a score of 150 on an SJT differed from a score of 200 from a practical and behavioral standpoint—an understanding of the relationship between scores and these meaningful outcomes is a necessary challenge to consider in advancing validity research surrounding instruments that measure professional competence.

When to account for participant characteristics. Yet another challenge in this research is the question of when to account for participant characteristics as important mediators or moderators of observed performance on instruments that measure components of professional competence. With regards to empathy, research shows that there can be substantial differences in

measurements based on gender and age (Gerdes, Segal, & Lietz, 2010; Jolliffe & Farrington, 2006; Renate et al., 2011). In health professions education, other factors such as training, work experience, and personal experiences can also influence measurements of these constructs (Fjortoft, Van Winkle, & Hojat, 2011; Hojat, 2007; Nunes, Williams, Sa, & Stevenson, 2011). Overall, these factors are not often rigorously evaluated and there is limited evidence to understand how participant characteristics can account for variance in examinees' response processes or in their scores, and whether the variance should be accommodated. Of note, this research did not include a thorough investigation of the relevance of these factors that were collected and only considered how differences in work experience (i.e., students compared to practicing pharmacists) may relate to differences in performance. Greater attention should be paid to collecting information about participants and investigating when these factors should be addressed.

Limited interpretations when using a single assessment strategy. The last challenge identified was that there are inherent limitations when using an individual assessment strategy to measure components of professional competence. As previously outlined, the domain of professional competence integrates multiple constructs that are highly related and difficult to distinguish from other another. As a result, the use of a single assessment strategy is limited in the inferences that can be drawn about an individual's standing on a particular construct at a moment in time. Specifically, this research provided evidence to suggest that SJTs require participants to engage in complex decision-making processes; however, it is unclear if this assessment accurately accounts for all components of the construct being assessed as well as the behaviors and decision-making processes that may occur in practice. There is a need for multiple assessment strategies (e.g., interviews, observations, instruments, etc.) to offer a more substantial

evaluation of constructs related professional competence. Holistic strategies that integrate multiple assessment modalities should be investigated to determine if they yield a more comprehensive understanding of individuals. In addition, these strategies should be frequently repeated to consider how the standing on these constructs evolves and potential changes that can be expected as individuals develop over time or proceed through the curriculum.

Challenges with Research on Response Processes

This study also aimed to address a significant gap in SJT research regarding evidence of the response process. Research on response processes is an emerging field and as a result includes several challenges that were encountered during this study, which are outlined in this section.

Response process research as an emerging field. In general, research on response processes is a growing field in the literature as it has been highlighted as a critical component of validity evidence for assessments. The reason for this growth is due to increased understanding of how individuals learn, a greater emphasis on the importance of complex thinking, and advancements in data collection techniques that can be used to evaluate these processes (Ercikan & Pellegrino, 2017). Due the infancy of the field, however, this means that there are few recommendations for how to conduct high quality research on responses processes and the value of the research may not be readily perceived in the literature. Leighton (2017a), for example, states “verbal response data are still not considered obligatory for safeguarding the validity of inferences made about examinees based on their test scores” (p. 25). Throughout this research, it was difficult to identify which strategies would be optimal to evaluate the response process and often required the integration of approaches borrowed from other fields, such as survey development research. Overall, there was a need to ensure the rigor of the research was aligned

with standards expected in the measurement community, which was a significant challenge as there was few models to serve as guidance.

Response process research requires multiple methodologies. In addition, research on the response process requires in-depth analyses using multiple methodologies to better understand assessments of complex thinking (Ercikan & Oliveri, 2016; Nichols & Huff, 2017). Assessments that measure complex thinking are expected to activate cognitive response processes, which include moment-to-moment steps required to think and make decisions during the assessment (Pellegrino, Chudowsky, & Glaser, 2001). Investigating these cognitive processes requires data collection that forces participants to explicate their thoughts through think-aloud and cognitive interviews. These data are then analyzed using qualitative and quantitative approaches to describe the cognitive process based on the utterances shared. As a result, research in this area was challenging as it was resource intensive to conduct interviews, transcribe the conversations, code the transcripts, and draw conclusions from data. Moreover, other methodologies reported in the literature could have been considered but were not due to feasibility constraints; these strategies include response times, eye-tracking, and log data in electronic assessments (Oranje, Gorin, Jia, & Kerr, 2017). In summary, there are a host of methodologies that can be incorporated to evaluate response processes and researchers must balance the challenges of feasibility and rigor to ensure the research questions are adequately addressed and relevant validity data collected.

Response process research necessitates models, which can vary. Lastly, research on response processes presents another challenge in that it often necessitates models to describe the process. These models can be complex depending on the domain being assessed and they can vary depending on the groups being studied. Previous research in the field of response processes

has focused on domains such as math, science, language arts, and history—these domains often have very explicit steps in the cognitive process that can be modeled depending on the component being assessed (Leighton, 2017; Nichols & Huff, 2017). The challenge with this research is that models of non-cognitive constructs, such as empathy, are not well-developed and there is limited research to suggest that a similar decision-making process is used consistently. Moreover, these models could vary based on the groups being examined or other contextual factors presented in the scenario. For example, differences in gender, race, cultural experiences, or age of both the participants and the actors presented in the scenarios may influence the different cognitive processes that alter decision-making and thereby influence the final model. This can be influential in generating validity evidence based on response processes as the model developed may be highly sample dependent, but it may be treated as generalizable to a larger population. Additional research should include samples related to the examinees being tested (e.g., health professions students) as well as those that can confirm or contend the findings (e.g., non-health professions or service-field related students). Overall, the challenge for researchers to is clearly articulate the constraints of the model and to investigate how the model may differ across constructs and samples with varying characteristics.

Challenges with Designing and Conducting Research on SJTs

This study included the design, administration, and evaluation of an SJT intended to measure empathy among students and practicing pharmacists. As a result, there were significant challenges identified during this research regarding the design of and research on SJTs that are described within this section.

Resource intensive process. Similar to other high-stakes testing conducted in the professions, the design and administration of SJTs is a highly resource intensive process that

requires multiple checkpoints to ensure that an instrument is created to generate reliable and valid results (Davis-Becker & Muckle, 2017). When creating the SJT in this research study, there was no exception to this expectation. The process was challenging as the design of this SJT needed to address a complex construct (i.e., empathy) and required a panel of subject matter experts to create and evaluate potential test items that would also generate data to address the questions of the research study. In this case, resources included gathering pharmacy faculty and practicing pharmacists to write test questions during a brief workshop and for another group to then evaluate the questions. Schedule coordination, teaching, and optimizing the questions all took a significant amount of effort prior to administering the test.

The data collection and analysis process were also heavily resource intensive due to the qualitative focus of the study. Thirty interviews were conducted, each lasting approximately one-and-a-half hours; this was also combined with time required to transcribe and analyze the data amongst multiple researchers. In general, this study included a relatively small sample size and short instrument with only 12 questions—the resources would be expected to increase greatly as larger scale studies are considered, which is common for high-stakes assessments. In addition, SJT research often includes investigating the relationship of SJT performance to other variables such as personality assessments and other surveys that may incur costs for each administration (McDaniel, Hartman, Whetzel, & Grubb, 2007; Wolcott, Lupton-Smith, Cox, & McLaughlin, 2018). Overall, researchers should be aware that sufficient time, funds, and personnel must be allocated to generate a high-quality SJT that targets the designed construct and that can provide meaningful data for research purposes.

Awareness of contextual features that affect design and participant responses. The design and research of SJTs also presented a challenge as there has been minimal discussion in

the literature about the importance of contextual features such as the item setting, the actors included in the item, and the amount of details provided. This research aimed to address some of those questions, however, it was not well known prior to the study whether these factors would have a significant influence. Previous research has suggested the reducing item complexity in SJTs (e.g., small word counts) is best to minimize construct-irrelevant variance as more complex items correlate more so with tests of cognitive ability rather than the knowledge, skills, and abilities being measured (Clevenger, Pereira, Wiechmann, Schmitt, & Schmidt-Harvey, 2001; Weekly & Ployhart, 2005). It is recommended that test developers be provided with more explicit instructions about how items should be constructed with regards to these contextual features. The results of this study specifically did not include contextual features about the actors such as gender, race, age, and cultural backgrounds that also have the potential to influence the findings—test developers must be cognizant of the potential impact of these factors and the possibility of eliciting various biases that could consequently impact the response process. Published research should also begin to include greater details about the structure and details of SJT items to develop a better understanding of how these features may play a role in the observed results and participant performance.

Lastly, the findings of this study allude to potential challenges when creating SJTs that attempt to assess more than one construct (e.g., empathy, communication, integrity, etc.). As discussed, the ability to design SJT items that exclusively target one construct is a difficult as constructs can overlap and interact with features of the item such as the setting, the actors included in the items, and other contextual features. Moreover, the combination of the constructs being assessed may alter findings compared to the observed results using an SJT that measures those constructs individually. For example, the combination of evaluating integrity and empathy

within an SJT may lead to different responses compared to evaluating conflict management and empathy if there are varying degrees of overlap in terms of the subcomponents of each of the constructs. Designers, researchers, and users of SJTs must be aware of how these elements influence the response process and thereby affect response selection and SJT performance.

Variation in response formats and scoring strategies. Another challenge with SJT design and research is that there can be considerable variation in the response formats and scoring strategies that are used; these differences have been shown to alter reliability coefficients and correlations with other variables and these effects are not insignificant (Bergman et al., 2006; De Lang et al., 2017). In this research study, all items included the same response format (i.e., ranking) in an effort to encourage participants to describe how they evaluated each of the response options. There are challenges, however, in that SJTs can use other types of response formats that can affect the interpretation of the results and, thus, the findings cannot always be generalizable. The same is true of scoring strategies; SJTs often utilize partial credit scoring techniques and the number of points awarded can differ if there are penalties assigned for incorrect answers (Weekley, Ployhart, & Holtz, 2006; De Lan et al., 2017). Overall, it is imperative that researchers provide reasoning for the response formats and scoring strategies that are used and additional research is necessary to identify if certain approaches are preferred.

Lack of best practice recommendations. The range of difficulties in designing and conducting research on SJTs all relate to an overarching challenge—there is a lack of best practice recommendations regarding SJT design and research. Several review articles have been published to summarize the findings from SJT literature; these offer some general recommendations and options regarding design and research strategies (Lievens, Peeters, & Schollaert, 2008; McDaniel & Nguyen, 2001; Patterson, Zibarras, & Ashworth, 2016; Whetzel &

McDaniel, 2009). A significant limitation of this body of work, however, is that there is substantial variability in SJT content, formats, scoring, and administration that makes comparisons among the administered instruments difficult. As a result, the recommendations are not well-supported, and this leads to researchers engaging in diverse types of design and research approaches.

Furthermore, best practice recommendations will need to delignate how results, interpretations, and uses of SJTs may differ based on the specific construct or combination of constructs being assessed. As discussed earlier, the assessment of certain constructs such as integrity or adaptability may not include interaction effects with the setting or the experiences that are recalled in comparison to constructs such as conflict management or empathy. Guidelines may necessitate very specific recommendations that are exclusive to particular elements and will not be generalizable. In summary, greater emphasis on systematic research to compare design and research approaches is necessary in future work.

Limitations of the Present Research Design

In addition to the overarching challenges of conducting research on the response process related to an SJT that measured a component of professional competence, there were also specific limitations associated with this study.

Limitations of the research methodology. The presented research had several limitations due to the focus of the study and the methodologies utilized to address the research questions. First, when considering the breadth of validity evidence that is suggested to be collected when designing an instrument (AERA, APA, & NCME, 2014), the focus of this research was predominantly focused on only one of the suggested elements—the response process. Although additional analyses were conducted to provide evidence that this SJT

produced valid and reliable data; this was not a substantial focus of the research and is, therefore, limited in providing a comprehensive evaluation of the validity evidence for SJTs. Moreover, the research questions focused on addressing validity evidence focused exclusively on support for the interpretation of score meaning and did not include any evidence gathering to support the *use* of SJT scores in practice.

With regards to the research methodologies employed, the study was advanced in that it integrated rigorous interview strategies as well as qualitative and quantitative analyses to address the research question. There were inherent limitations, however, with the design of the interviews that were not identified until the study was initiated. The most prominent limitation was that order of the cognitive interview questions may have influenced participant responses, especially as it pertained to the question about what participants thought the item was measuring. The question prior to this was related to a change in the setting, which often included a summary of the test item and may have identified salient features of the item that participants would not have been aware of if they had not been asked that question initially. In addition, not supplying participants with a list of constructs made the identification process much more challenging and led to significant variation in their responses.

Lastly, although the research study included the collection of substantial amounts of qualitative data, coding was used to quantify utterances to allow for quantitative comparisons more readily. This approach is common when using think-aloud and cognitive interviews, however, this deviates from the traditional paradigms of qualitative data analysis (Leighton, 2017a; Merriam & Tisdell, 2016; Willis, 2015). In addition, quantifying utterances has the potential to artificially inflate or deflate the prevalence of codes depending on how much a participant spoke. For example, one person may speak for a prolonged period and, therefore,

increase frequency counts for a code that may not occur very prominently for other participants that did not speak as much. The only strategy used to mitigate this was to limit the presence of codes to once per turn; however, other strategies to account for differences based on total speaking time may have been more optimal.

Limitations of SJT content and format. Another significant limitation of the study was in with regards to content and format of SJTs. This SJT was focused exclusively on one construct (i.e., empathy) and one practice setting (i.e., pharmacy). Moreover, most of the questions focused on situations more likely to occur in an inpatient or hospital pharmacy setting compared to other pharmacy practice setting such as community, ambulatory care, or industry. Consequently, the findings may not be generalizable to other health professions or SJTs that measure other constructs.

In addition, this SJT used a ranking response format to promote greater discussion of the decision-making process by participants. This response format, however, is optimal for situations that require prioritization of tasks and may not be ideal for situations in which there are responses that are definitively inappropriate (Weekley, Ployhart, & Holtz, 2006). Therefore, there may have been several items in that study where the ranking response format was not the optimal response strategy, which could have adversely impacted the findings. Moreover, the ranking format assumes that the separation between the ranked items are equal; however, several participants stated that the interpretation of the ranking would differ depending on the other response options that were being compared. In other words, ranking a response option as a "3" does not necessarily have the same meaning across all test items.

Limitations of the participant sample. Lastly, there were significant limitations due to the participants sampled in the study. Participants for this study were recruited using

convenience sampling at one school of pharmacy and local hospitals. In addition, the sample was predominantly women and was not necessarily representative of the larger population of student pharmacists and practicing pharmacists. It is also possible the use of convenience sampling created a biased group of participants who were more motivated to participate or more likely to be in higher standing on the construct of interest, skewing the results and resulting in data that were not representative of the variation that would be expected in a larger sample or population.

Moreover, data collection about the participants was limited for feasibility purposes and did not include a thorough assessment of factors that may have influenced the study findings. For instance, the research did not include the collection of pertinent information about participant perceptions of empathy or their definition of empathy. It is possible when participants labeled items as measuring empathy, they may have been using a different definition that was not consistent with the one used to create the instrument. There are also generational differences that may have been present in how students and practicing pharmacists view the significance of empathy in patient care. Health professions curricula are integrating more training that addresses how empathy can be used to connect with patients and improve patient outcomes; these training practices were not often included in previous curricula and may be a significant influence in performance (Hojat, 2007; Quince, Tiemann, Benson, & Hyde, 2016). In addition, the study did not substantially investigate how person-level characteristics related to SJT performance and could not account for variance that was attributable to these variables.

Future Research

Throughout the chapter, areas of future research were identified as they pertained to the specific research question and challenges encountered in this study. This section includes

additional areas of research to be considered as it relates to improving the body of literature around the validity evidence for SJTs.

Modifications to the research design of this study. The limitations of this study could be addressed in future research in at least five ways. First, it would be beneficial to expand the number of questions included in this SJT to increase the potential for providing a more reliable measure of the construct of interest. Second, the development phase of SJT items should be prolonged and integrate a larger number of subject matter experts, especially those with a background in assessing empathy in the health professions. The goal would be to improve the quality of the items and ensure they are as aligned as possible with the construct of interest to minimize construct-irrelevant variance. Third, this SJT should be modified to provide different response formats that are matched with the potential response options based on the situations presented in individual items. For example, ranking responses should be used when prioritization is necessary whereas selecting the best options may be used when there are definitively inappropriate responses that should be identified by the examinee.

Fourth, additional participant characteristics should be collected to have a better understanding of their definition of empathy, the extent of their training, and their perceptions about the significance of empathy in patient care. Moreover, additional instruments should be administered to participants such as those assessing their personality traits, proclivity for social desirability, and instruments that measure empathy and other pertinent constructs. The relationship of SJT performance to these measures can provide additional validity evidence to support the findings. Lastly, the cognitive interviews should be modified to minimize order effects of the questions that may bias participant response. This could include shuffling the questions to be asked during the interview or distributing the desired questions across

participants. In addition, a list of possible constructs could be provided to participants to aid in their identification of the construct being assessed to determine if there is an appreciable impact.

Confirmation and evaluation of SJT response process models. Another direction for future research that should be prioritized are studies that confirm the response process model presented in this study to determine if the findings are reproducible. In addition, this model should be evaluated with SJTs that measure different constructs, engage other professions, and include various settings. Such research could identify whether the model is generalizable across SJTs in different domains or if it is domain-specific. Moreover, there was evidence to suggest that some components of the model may not be as pertinent due to the limited frequency that participants discussed features such as ability, general knowledge, and impression management. Further research is necessary to determine if these features should be excluded from the model or if they are context-dependent; for example, these features may be more prominent in a high-stakes testing environment compared to the study conducted. In addition, the relationship between the multiple components in the model and their relationship to overall performance on an SJT would be a critical area of research. This may be particularly significant as it relates to the strategies used during SJTs to generate a response; if there are certain strategies that are linked to better performance on SJTs, those could potentially be learned by examinees and influence the validity of the results. Lastly, it would also be beneficial to integrate participants in SJT research studies who are outside of the target population to determine how much of participant performance is related to job-specific knowledge and experiences compared to general knowledge and experiences.

Connection of SJT performance to observed behaviors. Another significant void in SJT literature is the relationship of SJT performance to observed behaviors in practice. Early

research on SJTs showed positive correlations with job performance evaluations (Campion, Ployhart, & MacKenzie Jr., 2014; Chan & Schmitt, 2002); however, this does not necessarily ensure that participants respond to SJTs similarly to how they would respond in real-life—overall, this greatly limits the inferences that can be made about examinees. Ideally, a study should be designed that presents participants with similar cases using an SJT and using a simulated interaction to determine how well responses selected on an SJT correlate with behaviors observed in practice. In addition, a multitrait-multimethod approach (Campbell & Fiske, 1959) could be used to create substantial validity evidence that examines how measurements of various constructs related to professional competence are related to one another using multiple assessment modalities.

Moreover, investigating the link between SJT performance and actual behaviors could advance an understanding of what cut-offs may be appropriate to indicate a "good" or "poor" amount of the construct of interest. For example, certain score cut-offs for assessments of empathy may be able to serve as a surrogate marker for behaviors that relate to positive patient outcomes in the health professions. If there is a greater understanding of how the standing on the construct relates to specific behaviors, this could significantly improve the interpretation of individual performance scores and support the use of these instruments more appreciably.

Evaluation of SJTs as longitudinal assessment strategies. An additional area of future research should focus on the potential for SJTs to serve as longitudinal assessment strategies in the health professions. Currently, SJTs are most often used in the admissions process (Patterson, Zibarras, & Ashworth, 2016) and it is unknown if SJTs can reliably measure changes in standing on a construct over time. SJTs have the potential to serve as instrument that may be able to document learner progress throughout a curriculum, identify those that may need remediation, or

to measure the impact of programming such as professional development targeting pertinent skill sets. Using SJTs as a longitudinal assessment would also allow researchers to investigate the impact of other variables on individual's trajectories such as work place culture, age, gender, or work experiences. Additional research should include the formulation of validity evidence to determine if SJTs can be used across these settings as longitudinal assessments.

Evaluation of SJT design features that impact performance. Lastly, additional research is warranted to better understand how the myriad of SJT design features can impact participant performance. An advantage of SJTs is that they are a versatile assessment strategy that can be adapted based on the purpose of the assessment, the setting, or the constructs being assessed to provide the best fit according to the need (Weekley & Ployhart, 2006). It is unclear, however, if certain design features are more likely to produce more reliable or valid results. Currently, there is a lack of best practice recommendations to guide SJT development. Therefore, it would be a useful to conduct a series of studies that systematically integrated different design strategies to evaluate the impact on performance and potential consequences for the validity evidence toward the goal of formulating a set of best practices. This research should also include investigations of how different design features may affect the fairness of an SJT as it pertains to groups that differ based on age, gender, race, socioeconomic status, and other pertinent characteristics.

Summary

The purpose of this study was to explore the response process of examinees as they completed an SJT intended to measure empathy of healthcare providers with varying levels of experience. This included: (1) identifying the salient factors and strategies of the response process, (2) evaluating the extent to which job-specific and general knowledge and experience

influence the response process, (3) determining the extent to which the item setting influences the response process, and (4) exploring the relationship of the examinee's ability to identify the construct being assessed with their performance. A sample of 30 participants (15 student pharmacists and 15 practicing pharmacists) completed an SJT designed to assess empathy. Each participant engaged in a think-aloud interview while they completed an SJT followed by a cognitive interview to better understanding their response process. The interviews were analyzed to address the four research questions posed in the study.

The results of this study indicate that SJT response processes can be described using a four-component process: comprehension, retrieval, judgment, and response selection. There is evidence to suggest there are multiple factors that influence each of these components to varying degrees in the response selection process. Most notably, there was evidence that job-specific knowledge and experience was more often referenced by participants, which suggests that SJTs more likely tap into job-specific information than general domain knowledge. In addition, the results showed that the setting of the item can have significant implications in the response process contrary to previous beliefs about SJTs. Lastly, there was inconclusive evidence to describe the role of the ability to identify the construct being assessed and the relationship with SJT performance. Overall, additional research is warranted to confirm these findings as this was the first study to offer a comprehensive investigation on the response process.

This final chapter also included a discussion of the challenges anticipated when measuring professional competency, evaluating response processes, and researching SJTs that should be considered in future research. Assessing components of professional competence, such as empathy, represents a significant challenge in research due to the overlap with other constructs of interest. Moreover, research on response processes for assessments is an emerging

field that often requires more qualitative and mixed research methodologies to address pertinent research questions. The research design integrated within this study was an example of the rigor that is necessary to address these needs. Lastly, future research should be used to expand the understanding of the response process for SJTs, especially as it relates to other professions both related and unrelated to healthcare as well as with regards to other constructs and response formats. The study was limited in scope due to the convenience sampling, the focus on only one construct measured with participants from one health profession and may have introduced bias because of the research process.

Despite the limitations of the present study and the challenges with studying professional competency, response processes, and SJTs, the results of this research made substantial contributions to SJT research. The results indicated that SJTs require participants to engage in complex decision-making process that integrate various features of their knowledge, experiences, and personal attributes. This study was the first to offer a comprehensive evaluation of the response process using rigorous qualitative methodologies and offers insights into a grossly under-researched field. In summary, it contributes to foundational steps necessary to generate validity evidence for SJTs to aid in score interpretation.

APPENDIX A. ITEM DEVELOPMENT SESSION HANDOUT

GOAL: Develop 24 situational judgment test items to be reviewed by a second group of subject matter experts

RESEARCH FOCUS: Understanding the knowledge, experiences, and strategies participants use to answer SJTs

- What **cognitive processes and strategies** are involved when examinees respond to SJT items?
- What is the **role of job-specific experiences** in the response process to SJT items?
- What is the **role of setting** presented in SJT items in the response process?
- What is the **role of the ability to identify the construct** being evaluated in the response process of SJT items?

DOMAIN DEFINITION: EMPATHY

- *Cognitive Empathy*: the ability to construct a working model of the emotional states of others
  - Perspective taking: intuitively putting oneself in another person's shoes in order to see things from his/her perspective
  - Online simulation: an effortful attempt to put oneself in another person's position by imagining what that person is feeling; likely related to future intentions
- *Affective Empathy*: the ability to be sensitive to and vicariously experience the feelings of others
  - Emotion contagion: the automatic mirroring of the feelings of others
  - Proximal responsivity: the affective response when witnessing the mood of others in a close social context
  - Peripheral responsivity: the affective response when witnessing the mood of others in a detached context

INSTRUCTIONS

- Each team will create six (6) test items to address one of the two areas of empathy
- Each item should have five (5) response options that will be ranked by the examinee
- Three (3) items should be in a healthcare setting
- Three (3) items should be in a non-healthcare setting (preferably not in a human services field)
- Each item should include a key that ranks options from most appropriate to least appropriate

TEAM ASSIGNMENTS

Team 1 & 2: affective empathy
Team 3 & 4: cognitive empathy

Cognitive Empathy	Affective Empathy
- See things from another person's point of view - Look at each side of a disagreement - Imagine someone's perspective / put myself in their shoes - When someone wants to enter a conversation - Predicting how someone will feel or what someone will do - When someone is feeling awkward or uncomfortable - Telling when someone is interested or bored when talking - Sense when you are intruding - Identify what a person wants to talk about - Know if someone is masking their true emotion - Consider when feeling upset or criticizing someone - Considering other people's feelings before acting	- Emotional during movies, films, or books - Get emotionally involved with friends' problems - Get nervous around others who feel nervous - People have a strong influence on your mood - Affected when a friend or someone close gets upset - Worried when others are panicking - Identify why things upset people - People talk to you because you're understanding

APPENDIX B. ITEM REVIEW SESSION HANDOUT

GOAL: Revise and create a key for 24 situational judgment test (SJT) items to be used in the final test

RESEARCH FOCUS: Understanding the knowledge, experiences, and strategies participants use to answer SJTs

- What **cognitive processes and strategies** are involved when examinees respond to SJT items?
- What is the **role of job-specific experiences** in the response process to SJT items?
- What is the **role of setting** presented in SJT items in the response process?
- What is the **role of the ability to identify the construct** being evaluated in the response process of SJT items?

DOMAIN DEFINITION: EMPATHY

- *Cognitive Empathy*: the ability to construct a working model of the emotional states of others
  - Perspective taking: intuitively putting oneself in another person's shoes in order to see things from his/her perspective
  - Online simulation: an effortful attempt to put oneself in another person's position by imagining what that person is feeling; likely related to future intentions
- *Affective Empathy*: the ability to be sensitive to and vicariously experience the feelings of others
  - Emotion contagion: the automatic mirroring of the feelings of others
  - Proximal responsivity: the affective response when witnessing the mood of others in a close social context
  - Peripheral responsivity: the affective response when witnessing the mood of others in a detached context

INSTRUCTIONS

- Each person will complete the SJT independently
- Rank each of the response options from most (1) to least (5) appropriate based on how you should respond
- Evaluate how well the test item measures empathy on a scale of 1 (very poorly) to 5 (very well)
- Identify if you think the question address affective (A) or cognitive (C) empathy
- Identify if you think the question includes a healthcare setting (Y) or not (N)
- Provide edits to improve the SJT questions

Cognitive Empathy	Affective Empathy
- See things from another person's point of view - Look at each side of a disagreement - Imagine someone's perspective / put myself in their shoes - When someone wants to enter a conversation - Predicting how someone will feel or what someone will do - When someone is feeling awkward or uncomfortable - Telling when someone is interested or bored when talking - Sense when you are intruding - Identify what a person wants to talk about - Know if someone is masking their true emotion - Consider when feeling upset or criticizing someone - Considering other people's feelings before acting	- Emotional during movies, films, or books - Get emotionally involved with friends' problems - Get nervous around others who feel nervous - People have a strong influence on your mood - Affected when a friend or someone close gets upset - Worried when others are panicking - Identify why things upset people - People talk to you because you're understanding

APPENDIX C. SITUATIONAL JUDGMENT TEST

ITEM: CH1

You notice a patient becoming upset with the physician during rounds. As the medical team begins to leave, the patient asks for you to stay behind. They tell you "I feel like the doctor never listens to me and they just do what they want without asking me first".

Rank each of the following response options based on how you SHOULD *respond to the scenario. Use 1 to indicate the MOST appropriate response and 5 to indicate the LEAST appropriate response. There can be no ties or duplicates.*

_____ Tell the physician what the patient said and suggest they address the patient's concerns.
_____ Ask the patient why they feel like they are not being heard.
_____ Tell the patient the doctor is "like this with everyone, it is nothing against you".
_____ Ask the nurse if the patient has been irritable recently or complaining when the team is not present.
_____ Tell the patient you understand how they feel and share a story about how you sometimes feel like people do not listen to you as well.

ITEM: CH2

When you contact a pharmacy to verify a patient's medication history, the pharmacist complains the store is really busy and that the information is probably "already in your system". The pharmacist asks you to call back later but the doctor is requesting the information before they start any new medications.

Rank each of the following response options based on how you SHOULD *respond to the scenario. Use 1 to indicate the MOST appropriate response and 5 to indicate the LEAST appropriate response. There can be no ties or duplicates.*

_____ Tell the pharmacist patient care is a top priority and this is essential information that will only take a few minutes to share.
_____ File a complaint about the pharmacist to the store's manager.
_____ Tell the pharmacist you understand how they feel and how hectic your job can be at times.
_____ Apologize for the inconvenience and convince the pharmacist of the urgency of the situation.
_____ Ask the pharmacist if there is an alternative or easier way to facilitate getting the information.

ITEM: CH3

According to your patient's blood sugar logs, he has always been within his goals; however, his other tests suggest his diabetes is poorly controlled and you suspect he has been recording false numbers.

Rank each of the following response options based on how you SHOULD *respond to the scenario. Use 1 to indicate the MOST appropriate response and 5 to indicate the LEAST appropriate response. There can be no ties or duplicates.*

_____ Ask how the patient has been taking his blood sugar and documenting his numbers.
_____ Ask the patient what problems he has with managing his diabetes, if any.
_____ Request he be transferred to a different pharmacist due to his lack of compliance.
_____ Tell the patient you suspect some of the numbers he provided may not be accurate based on the tests collected today.
_____ Contact a family member or caregiver to ask about his diabetes management at home.

APPENDIX C. SITUATIONAL JUDGMENT TEST (CONTINUED)

ITEM: CN1

You and a friend are studying for a big exam for one of your undergrad classes when they begin to complain about the course. Their parents have threatened to stop paying their tuition if they don't get an "A" in the course. Your friend tells you they purchased some medication off a friend who said it would help them study and they offer you some.

Rank each of the following response options based on how you SHOULD *respond to the scenario. Use 1 to indicate the MOST appropriate response and 5 to indicate the LEAST appropriate response. There can be no ties or duplicates.*

_____ Offer to help your friend find alternative ways to cope with the stress of the course and family.
_____ Tell your friend no, you don't need that to study.
_____ State that you will just take one to see what it does.
_____ Ask your friend if it's necessary to take the medication and whether that is a good idea.
_____ Acknowledge your friend's situation and discuss with them the challenges they are facing in and outside of the course.

ITEM: CN2

You go to the store to pick up a few things you forgot for a presentation. While standing in line at checkout, someone approaches you and asks if they can cut in front of you. However, there are already 5 people behind you. They mention that their children are at home sick and they are trying to get back as quickly as possible. Letting the person go in front of you will definitely make you late for your presentation.

Rank each of the following response options based on how you SHOULD *respond to the scenario. Use 1 to indicate the MOST appropriate response and 5 to indicate the LEAST appropriate response. There can be no ties or duplicates.*

_____ Ask the people behind you if they would mind having the person go in front of you.
_____ Acknowledge their situation and let them go in front of you.
_____ Tell them no and that they need to get in line like all the others.
_____ Ask the person what is wrong with their children and determine whether they cut can based on their response.
_____ Tell them that you are also in a rush and ask if they could cut in front of the person behind you.

ITEM: CN3

You are having dinner with several family members when your parents start asking about your sibling's marital status. They ask your sibling a series of questions: what is going on, why they haven't been successful, and other details. Your sibling begins to look uncomfortable with the questions.

Rank each of the following response options based on how you SHOULD *respond to the scenario. Use 1 to indicate the MOST appropriate response and 5 to indicate the LEAST appropriate response. There can be no ties or duplicates.*

_____ Ask family members if they could share their own answers or challenges to the questions they are asking.
_____ Respectfully divert conversation to a new topic of discussion.
_____ Tell your family members to back off on questioning and that it is not their business.
_____ Join in on the questioning to make your sibling respond.
_____ Acknowledge to family members that these are difficult questions and may not be the best time to discuss.

APPENDIX C. SITUATIONAL JUDGMENT TEST (CONTINUED)

ITEM: AH1

One of your patients appears to be very depressed, which they believe to have been precipitated by the recent loss of a loved one. You realize their loss parallels one of your own experiences and wonder how this might be used to develop rapport with your patient.

Rank each of the following response options based on how you SHOULD *respond to the scenario. Use 1 to indicate the MOST appropriate response and 5 to indicate the LEAST appropriate response. There can be no ties or duplicates.*

_____ Describe your own loss and subsequent feelings.
_____ Acknowledge their understandable sadness from experiencing a personal loss.
_____ Change the subject, as dwelling on it may make them more upset.
_____ Encourage the patient to discuss their feelings with a friend, family member, or religious leader.
_____ Recommend they speak more with their provider about counseling services.

ITEM: AH2

A nurse interrupts you during rounds about a patient on the floor not covered by your service. They tell you a family member noticed that the infusion rate for a medication was incorrect on the pump. The nurse has asked you to talk with them. When you enter the room the several family members are very upset with the situation.

Rank each of the following response options based on how you SHOULD *respond to the scenario. Use 1 to indicate the MOST appropriate response and 5 to indicate the LEAST appropriate response. There can be no ties or duplicates.*

_____ Apologize to the patient and the family about the error.
_____ Discuss with the patient and family about the potential complications from infusing the medication at the wrong dose.
_____ Leave the room to allow the patient and family to process what has occurred because the medication error was not directly your fault.
_____ Ask the patient and family what questions they have about the medication error.
_____ Identify the potential causes of the error and consider ways to avoid the error in the future.

ITEM: AH3

You were asked by a physician to speak with a patient's family about the upcoming chemotherapy treatment for their 8-year old son. When you start talking about the negative side effects of the drug treatment, the patient's parent becomes visibly upset and asks you to "stop talking about this."

Rank each of the following response options based on how you SHOULD *respond to the scenario. Use 1 to indicate the MOST appropriate response and 5 to indicate the LEAST appropriate response. There can be no ties or duplicates.*

_____ Tell that patient's parent it is hospital policy to review all of the necessary information before beginning chemotherapy and you are required to finish.
_____ Tell the physician the family refused to complete the education and became upset.
_____ Conclude the session and document education has been complete.
_____ Request to schedule a different time to continue discussing the medication when the family would be more comfortable.
_____ Ask the parent about their concerns with the medication.

APPENDIX C. SITUATIONAL JUDGMENT TEST (CONTINUED)

ITEM: AN1

You are shopping at a grocery store with one of your parents when they start to behave strangely. You parent starts to get very anxious about someone else they saw in the store. They keep saying it would be best to leave and come back at a later time.

Rank each of the following response options based on how you SHOULD *respond to the scenario. Use 1 to indicate the MOST appropriate response and 5 to indicate the LEAST appropriate response. There can be no ties or duplicates.*

_____ Become concerned and ask what the issue is.
_____ Tell your parent not to worry and they are just imagining things.
_____ Suggest to your parent that you could confront the other person.
_____ Leave the store immediately with your parent.
_____ Comfort your parent but continue shopping to complete your errand.

ITEM: AN2

One of your closest relatives has been trying desperately to conceive a child over the past few years with no success. During lunch one day, they describe their frustrations and begin to become visibly upset. Your relative talks about how they feel responsible for the issue and feel like "so much is out of [their] control".

Rank each of the following response options based on how you SHOULD *respond to the scenario. Use 1 to indicate the MOST appropriate response and 5 to indicate the LEAST appropriate response. There can be no ties or duplicates.*

_____ Provide comfort to your relative and ask further questions about their feelings.
_____ Discuss if they've considered other options such as fertility clinics, surrogacy, or adoption.
_____ Offer to arrange a time to speak with them again to follow up if things have improved.
_____ Comfort them by talking about how difficult it is to raise children and how it limits lifestyle.
_____ Describe your own struggles with becoming pregnant or stories of other couples.

ITEM: AN3

One of your best friends visits you during college. One evening during dinner they begin to tell you that they are planning to drop out of school. They begin to list the factors they have weighed while making their decision and begin to cry. They say there are too many "overwhelming obstacles" and they are "not cut out for college".

Rank each of the following response options based on how you SHOULD *respond to the scenario. Use 1 to indicate the MOST appropriate response and 5 to indicate the LEAST appropriate response. There can be no ties or duplicates.*

_____ Offer a hug or a moment to let them reflect on the discussion because they are crying.
_____ Request they list the issues they've encountered and offer strategies you've used before to prioritize managing college life.
_____ Acknowledge that feeling overwhelmed is a common occurrence in college.
_____ Ask if your friend has sought support from school administrators or talked to anyone about it if sought support and/or talked to anyone.
_____ Provide support for their decision and acknowledge that decision was difficult.

APPENDIX D. RECRUITMENT EMAILS

Student and Pharmacist Recruitment Email

Dear [insert name],

The UNC Eshelman School of Pharmacy is conducting a study to describe how examinees respond to an instrument used in the health professions: the *situational judgment test* (*SJT*). The SJT includes a series of cases and asks examinees to evaluate which response would be most appropriate. The goal is to better understand what information is used when respondents answer test questions.

As part of this study, you will be asked to participate in a one-on-one interview for 1 to 1.5 hours. Based on your availability, we will coordinate a time that works best with your schedule. The interview may be conducted via videoconference (e.g. Zoom™) to minimize the necessity for travel. During this timeframe, we will ask you to answer the test questions and describe your thought process in selecting the best answer. You will not need to prepare in advance.

Please complete this survey to indicate your interest:
https://unc.az1.qualtrics.com/jfe/form/SV_3DyHQ4mHqexCe9v

If you have any questions about the study, please contact Michael Wolcott at wolcottm@email.unc.edu.

Thank you in advance for your consideration of this request.

Sincerely,

Michael Wolcott

Michael Wolcott, PharmD, BCPS
PhD Candidate, Learning Sciences and Psychological Studies
University of North Carolina School of Education

Graduate Research Assistant
UNC Eshelman School of Pharmacy
321 Beard Hall, Chapel Hill, NC 27516
Phone: 919-451-3547

APPENDIX E. PARTICIPANT CONSENT DOCUMENT

University of North Carolina at Chapel Hill
Consent to Participate in a Research Study
Interview for Adult Participants

Title of Study: Describing the response process during a situational judgment test
Principal Investigator: Michael Wolcott
Principal Investigator Department: UNC Eshelman School of Pharmacy
Principal Investigator Phone number: (919) 451-3547
Principal Investigator Email Address: wolcottm@email.unc.edu
Faculty Advisor: Jacqui McLaughlin
Faculty Advisor Contact Information: (919) 966-4557

What are some general things you should know about research studies?
You are being asked to take part in a research study. To join the study is voluntary. You may choose not to participate, or you may withdraw your consent to be in the study, for any reason, without penalty.

Research studies are designed to obtain new knowledge. This new information may help people in the future. You may not receive any direct benefit from being in the research study. There also may be risks to being in research studies.

Details about this study are discussed below. It is important that you understand this information so that you can make an informed choice about being in this research study.

You will be given a copy of this consent form. You should ask the researchers named above, or staff members who may assist them, any questions you have about this study at any time.

What is the purpose of this study?
The purpose of this research study is to conduct a comprehensive evaluation of the response process examinees use when completing a situational judgment test. You are being asked to be in this study because you are a student or a practicing pharmacist.

Are there any reasons you should not be in this study?
You should not be in this study if you feel you cannot complete the situational judgment test as intended or if you feel you are not able to share your thoughts about the response process.

How many people will take part in this study?
There will be approximately 40 people in this research study.

How long will your part in this study last?
Your participation in this interview will last approximately one and one-half hours.

What will happen if you take part in the study?
You will be asked to answer situational judgment test items and to describe your thought process in selecting your answers. You will then be asked to answer additional questions to determine other factors that contributed to your selections. You may choose to respond or not respond at any point during the discussion. The interview will be audio or video-recorded so we can convert the interview to a transcript.

What are the possible benefits from being in this study?
We do not anticipate direct benefits to you as a participant in the study. The findings are anticipated to benefit pharmacy practice as a whole, which may indirectly offer benefits to your experiences as a clinician or student.

APPENDIX E. PARTICIPANT CONSENT DOCUMENT (CONTINUED)

What are the possible risks or discomforts involved from being in this study?
We do not anticipate any risks or discomfort to you from being in this study. All data collected will be confidential; therefore, we encourage you to be as honest and open as you can.

How will information about you be protected?
Every effort will be taken to protect your identity as a participant in this study. You will not be identified in any report or publication of this study or its results. Your name will not appear on any transcripts; instead, you will be given a code number or pseudonym. The list which matches names and code numbers / pseudonyms will be kept in a locked file cabinet. After audio- or video-recordings have been transcribed, the recording will be destroyed, and the list of names and numbers will also be destroyed.

What if you want to stop before your part in the study is complete?
You can withdraw from this study at any time, without penalty. The investigators also have the right to stop your participation at any time. This could be because you have had an unexpected reaction, or have failed to follow instructions, or because the entire study has been stopped.

Will you receive anything for being in this study?
You may receive compensation for participating in the study, including a gift card or food.

What if you are a UNC student?
You may choose not to be in the study or to stop being in the study before it is over at any time. This will not affect your class standing or grades at UNC-Chapel Hill. You will not be offered or receive any special consideration if you take part in this research.

What if you are a UNC employee?
Taking part in this research is not a part of your University duties, and refusing will not affect your job. You will not be offered or receive any special job-related consideration if you take part in this research.

What if you have questions about this study?
You have the right to ask, and have answered, any questions you may have about this research. If you have questions about the study (including payments), complaints, concerns, or if a research-related injury occurs, you should contact the researchers listed on the first page of this form.

What if you have questions about your rights as a research participant?
All research on human volunteers is reviewed by a committee that works to protect your rights and welfare. If you have questions or concerns about your rights as a research subject, or if you would like to obtain information or offer input, you may contact the Institutional Review Board at 919-966-3113 or by email to IRB_subjects@unc.edu.

Participant's Agreement:
I have read the information provided above. I have asked all the questions I have at this time. I voluntarily agree to participate in this research study.

_____ _____
Signature of Research Participant Date

_____
Printed Name of Research Participant

_____ _____
Signature of Research Team Member Obtaining Consent Date

_____
Printed Name of Research Team Member Obtaining Consent

APPENDIX F. SJT ITEM AND INTERVIEW DISTRIBUTION PER PARTICIPANT

Participant	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12
P01	AN3	AN1	CN3	AN2	CN1	AH3	CH1	AH1	CH2	CN2	AH2	CH3
P02	AN3	CH1	CH2	AH2	AN1	AH3	CN1	CN3	AH1	CN2	AN2	CH3
P03	CH3	CN2	AN3	CH1	AH1	AH2	CN3	AN2	CH2	CN1	AH3	AN1
P04	AH1	AN1	CN3	CH2	CH3	CN1	CN2	AN2	AH3	AN3	AH2	CH1
P05	AN3	AH1	AN1	CH1	AH2	AN2	CN2	CN1	AH3	CN3	CH2	CH3
P06	AH3	AN3	CN2	AH1	CH1	CN1	AN1	CH2	AH2	AN2	CN3	CH3
P07	CN2	CH1	CH2	AH1	AN3	CN1	AN2	AH3	AH2	CN3	AN1	CH3
P08	CN3	AH3	CH3	CH1	AN3	CH2	AN1	AH1	CN1	AN2	CN2	AH2
P09	CN3	AH2	CH1	CN1	AN2	AH3	CN2	CH3	AN3	CH2	AN1	AH1
P10	AN2	CN2	CH3	AH3	AN1	CN1	CN3	AN3	CH1	AH2	AH1	CH2
P11	AN2	AH3	CH3	AN3	AH2	AN1	CN1	CH1	AH1	CN3	CH2	CN2
P12	AN2	CN3	CH1	CH2	CN1	AH1	AN1	CN2	AH2	AN3	CH3	AH3
P13	AH3	CN1	CH3	CH2	CN2	AN1	AH2	AN3	CN3	AH1	CH1	AN2
P14	CH1	AH2	CH3	AN3	CN1	CN3	AN1	AN2	AH3	CH2	CN2	AH1
P15	CH3	CH2	AH3	AN2	AN1	CH1	CN1	CN3	AH2	AH1	AN3	CN2
S01	CH2	CH1	AH2	AN2	AN3	AN1	AH3	CN1	AH1	CH3	CN3	CN2
S02	CH1	AN3	AN2	AH3	AH1	CN1	AH2	CN3	AN1	CH2	CN2	CH3
S03	CH1	CN3	CN1	CH3	CN2	AH2	AN1	AN3	AH3	AH1	CH2	AN2
S04	CN1	AN2	AN1	CN2	AN3	CN3	AH1	CH2	AH3	AH2	CH3	CH1
S05	CN3	AH3	AN2	CH2	CH3	AH1	CN1	CN2	CH1	AH2	AN1	AN3
S06	AH1	AN3	CH2	AN2	CN2	CN3	CH1	CN1	CH3	AH2	AN1	AH3
S07	AN2	AN1	CH3	CH1	AH2	AN3	CN2	CN1	CH2	CN3	AH1	AH3
S08	CH2	AN3	CH3	CH1	AH2	AH3	AN1	CN3	CN2	CN1	AN2	AH1
S09	CN1	CH2	CH1	AN3	CN2	AH3	CH3	AN1	AH2	CN3	AN2	AH1
S10	AH1	CH2	CN2	CH1	AN3	CN3	AN2	AH2	AH3	AN1	CH3	CN1
S11	CH1	AH1	AH3	CN3	CH3	CN2	AH2	AN1	AN2	CH2	CN1	AN3
S12	AH2	CN2	CH2	CH1	AN1	AN2	AH3	AH1	CN3	CH3	CN1	AN3
S13	CN1	AN2	AH3	AN1	CN2	AH2	CH2	CH1	CN3	AN3	AH1	CH3
S14	AN1	AN2	CN1	AH3	CH1	AH1	CH3	AN3	CN2	AH2	CH2	CN3
S15	AH1	AN1	CH2	AH3	CN2	AN3	AH2	CH1	CN1	CH3	AN2	CN3

EXTRA												
E01	11	5	12	7	4	8	6	1	9	3	10	2
E02	2	4	5	6	8	11	3	9	10	1	12	7
E03	5	11	8	2	10	9	6	3	7	12	1	4
E04	4	2	11	12	3	1	6	10	7	9	8	5
E05	3	10	6	4	1	9	5	7	8	2	12	11

		COUNT
CH1	1	30
CH2	2	30
CH3	3	30
CN1	4	30
CN2	5	30
CN3	6	30
AH1	7	30
AH2	8	30
AH3	9	30
AN1	10	30
AN2	11	30
AN3	12	30

CH	CN	AH	AN
1,2	2,3	1,3	1,3
1,3	1,3	2,3	1,3
1,3	1,2	1,3	2,3
2,3	1,2	2,3	2,3
2,3	2,3	2,3	2,3

PHARMACIST					
COUNT(1)	10	10	10	10	40
COUNT(2)	10	10	10	10	40
COUNT(3)	10	10	10	10	40

STUDENT					
COUNT(1)	10	10	10	10	40
COUNT(2)	10	10	10	10	40
COUNT(3)	10	10	10	10	40

CH	CN	AH	AN
1,2	2,3	1,2	2,3
1,3	1,3	1,3	1,3
2,3	2,3	1,2	2,3
1,2	1,2	1,3	1,2
1,2	1,2	1,3	1,2
2,3	2,3	1,3	2,3
1,3	1,3	2,3	2,3
1,2	1,2	1,3	1,2
1,3	1,2	2,3	1,2
1,3	2,3	2,3	1,3
2,3	2,3	1,2	1,2
2,3	1,2	1,2	1,3
1,2	1,3	2,3	2,3
1,3	1,3	2,3	1,3
2,3	1,3	1,2	1,3
1,3	1,2	1,3	1,2
1,2	2,3	2,3	2,3
2,3	1,2	1,3	1,3
1,3	1,3	1,2	1,3
1,2	1,3	2,3	2,3
2,3	1,3	2,3	1,3
2,3	2,3	2,3	1,2
1,2	2,3	1,2	1,2
1,2	2,3	1,2	1,3
2,3	1,2	1,3	1,2
2,3	1,3	2,3	2,3
1,3	1,2	1,2	2,3
1,3	2,3	1,3	2,3
1,2	1,2	1,2	1,3
1,3	1,3	1,3	1,2

APPENDIX G. QUESTIONNAIRE OF COGNITIVE AND AFFECTIVE EMPATHY

Please rate your level of agreement with the following statements on a scale of 1 = Strongly Disagree to 4 = Strongly Agree
[**MARK YOUR RESPONSE WITH A ✓]**

Item	Statement	Strongly Disagree	Disagree	Agree	Strongly Agree
1	I sometimes find it difficult to see things from the "other guy's" point of view.				
2	I am usually objective when I watch a film or play, and I don't often get completely caught up in it.				
3	I try to look at everybody's side of a disagreement before I make a decision.				
4	I sometimes try to understand my friends better by imagining how things look from their perspective.				
5	When I am upset at someone, I usually ty to "put myself in his shoes" for a while.				
6	Before criticizing somebody, I try to imagine how I would feel if I was in their place.				
7	I often get emotionally involved with my friends' problems.				
8	I am inclined to get nervous when others around me seem to be nervous.				
9	People I am with have a strong influence on my mood.				
10	It affects me very much when one of my friends seems upset.				
11	I often get deeply involved with the feelings of a character in a film, play, or novel.				
12	I get very upset when I see someone cry.				
13	I am happy when I am with a cheerful group and sad when the others are glum.				
14	It worries me when others are worrying and panicky.				
15	I can easily tell if someone else wants to enter a conversation.				
16	I can pick up quickly if someone says one thing but means another.				
17	It is hard for me to see why some things upset people so much.				
18	I find it easy to put myself in somebody else's shoes.				
19	I am good at predicting how someone will feel.				
20	I am quick to spot when someone in a group is feeling awkward or uncomfortable.				
21	Other people tell me I am good at understanding how they are feeling and what they are thinking.				
22	I can easily tell if someone else is interested or bored with what I am saying.				
23	Friends talk to me about their problems as they say that I am very understanding.				
24	I can sense if I am intruding, even if the other person does not tell me.				
25	I can easily work out what another person might want to talk about.				
26	I can tell if someone is masking their true emotion.				
27	I am good at predicting what someone will do.				
28	I can usually appreciate the other person's viewpoint, even if I do not agree with it.				
29	I usually stay emotionally detached when watching a film.				
30	I always try to consider the other fellow's feelings before I do something.				
31	Before I do something I try to consider how my friends will react to it.				

APPENDIX H. STUDENT PARTICIPANT DEMOGRAPHIC SURVEY

Please complete the following questionnaire to describe your background and experiences.

What is your current age? _____ ☐ I prefer not to respond

Which of the following best describes your identified gender?

☐ Male ☐ Female ☐ I prefer not to respond

Which of the following best describes your pharmacy school status?

☐ First-Year Student Pharmacist (c/o 2022)
☐ Second-Year Student Pharmacist (c/o 2021)
☐ Third-Year Student Pharmacist (c/o 2020)
☐ Fourth-Year Student Pharmacist (c/o 2019)

Which of the following best describes your education status (<u>select all that apply</u>)?

☐ Bachelor of Science Degree (major: _____)
☐ Bachelor of Arts Degree (major: _____)
☐ Master's Degree (major: _____)
☐ Doctoral Degree (major: _____)

What work experience do you have in the health professions?

What is the average <u>number of hours</u> you work per week in a healthcare setting? _____

What is the average <u>number of patients</u> you interact with on a weekly basis in a healthcare setting? _____

What is the average <u>number of healthcare providers</u> (*NOT* including pharmacists) you interact with on a weekly basis? _____

What work experience do you have in other human services-related fields (e.g. retail, food services, etc.)?

How many <u>years</u> of work experience do you have in <u>other</u> human services-related fields (e.g. retail, food services, etc.) _____

Do you have experience taking care of a family member or individual who was terminally ill? (circle response) Yes No

How do you define "empathy"?

What type of empathy training have you completed (e.g. readings, workshops, evaluations, etc.)?

How important is empathy to your future work as a pharmacist? Why?

APPENDIX I. PHARMACIST PARTICIPANT DEMOGRAPHIC SURVEY

Please complete the following questionnaire to describe your background and experiences.

What is your current age? _____ ☐ I prefer not to respond

Which of the following best describes your identified gender?

☐ Male ☐ Female ☐ I prefer not to respond

Please select all of the following education training / certifications you have completed (do NOT include activities in process):

☐ Doctor of Pharmacy (PharmD)
☐ Advanced Degree (MPH, MBA, PhD, etc.)
☐ First-Year Residency Program (PGY1)
☐ Fellowship or Post-Doctoral Position
☐ Second-Year Residency Program (PGY2)
☐ Board of Pharmacy Specialties Certification (e.g. BCPS, etc.)

What do you identify as your primary practice area (e.g. current or previous practice)?

☐ Administration ☐ Critical Care / Emergency Medicine ☐ Oncology
☐ Ambulatory Care ☐ Drug Information General Medicine ☐ Pediatrics
☐ Cardiology ☐ General Surgery ☐ Transplant
☐ Community ☐ Neurology ☐ Other- please describe _____

Where is / was your primary practice area located (if you no longer practice, select where you current reside)?

☐ Research Triangle Park (e.g. Raleigh, Durham, or Chapel Hill)
☐ Western North Carolina (e.g. Asheville)
☐ Central North Carolina (e.g. Greensboro)
☐ South Western North Carolina (e.g. Charlotte)
☐ Eastern North Carolina (e.g. Fayetteville, Elizabeth City, Greenville)

What year were you first licensed as a pharmacist (including outside of North Carolina)? _____

How many <u>years</u> have you had a faculty appointment at a health professions school? _____

What is the average <u>number of hours</u> you work per week in a healthcare setting? _____

What is the average <u>number of patients</u> you interact with on a weekly basis in a healthcare setting? _____

What is the average <u>number of students</u> (of any health profession) you interact with on a weekly basis in a healthcare setting?_____

What is the average <u>number of healthcare providers</u> (*NOT* including pharmacists) you interact with on a weekly basis? _____

What work experience do you have in other human services-related fields (e.g. retail, food services, etc.)?

How many <u>years</u> of work experience do you have in <u>other</u> human services-related fields (e.g. retail, food services, etc.) _____

Do you have experience taking care of a family member or individual who was terminally ill? (circle response) Yes No

How do you define *empathy*?

What type of empathy training have you completed (e.g. readings, workshops, evaluations, etc.)?

How important is empathy to your work as a pharmacist? Why?

APPENDIX J. THINK-ALOUD INTERVIEW SCRIPT

The following script was adopted from Leighton (2017).

Thank you for attending the session today.

Today's session will be divided into two parts. In the first part, you will complete twelve (12) questions on a fictitious exam that could be used to evaluate potential residents for a residency program or students for a health professions program. For each question, you will be given a scenario and requested to rank the response options based on how you *should* respond to the scenario. Your rankings should be labeled 1 for the most appropriate and 5 for the least appropriate with no ties or duplicates. In the second part, I will be asking you specific questions about a randomly selected set of eight (8) questions.

For the first part of this study, I am interested in learning about the thoughts you have as you answer. For this reason, I am going to ask you to think aloud as you work through the test. Let me explain what I mean by "think aloud". It means that I would like you to tell me everything you think about as you work through each test question. You will do this one test question at a time.

When I say tell me everything, I really mean every thought you have from the moment you read the problem to the end when you have a solution or even if you do not have a solution. Please do not worry about planning how to say things or clarifying your thoughts. What I really want is to hear your thoughts constantly as you try to solve the problem – uninterrupted and unedited. Sometimes you may need time to think quietly through something – if so, this is okay but please tell me what you thought through as soon as possible after you are finished.

I realize it can feel awkward to think aloud but try to imagine you are alone in the room. If you become silent for too long, I will say "keep talking" to remind you to think aloud. Please note, this research is highly exploratory. My intention is not to evaluate your thinking or explanations while you speak. The purpose of the study is to learn about the thoughts as you—and other people—answer each question.

We will have an opportunity to practice, but before we get to that, please let me know if you understand what we will be doing today.

Do you have any questions?

Let us now practice thinking aloud with two practice problems presented on your paper.

- Lucas works 7.5 hours in a day. How many hours does he work in 5 days? *Now, please tell me everything that you are thinking as you try to solve this*.
- What is the 5th letter after C? *Now, please tell me everything that you are thinking as you try to solve this*.

APPENDIX K. COGNITIVE INTERVIEW SCRIPT

The following script was adopted from Leighton (2017).

Begin the interview and start with the first selected test question – after the participant reviews each question, the interviewer will ask the following if it was not addressed by the participant:

For this next part, I will ask you a series of questions about each question – they will become repetitive. Please be succinct in your responses. At the end, I will ask some general questions about the test as a whole.

Do you have any questions?

- How did you decide how to rank each option?
    - *Further probe:* What made your decisions easier and why?
    - *Further probe:* What made your decisions harder and why?

- What, if any, experiences does this question make you think of when you answered the question?
    - *Further probe:* What memories did you think about when you answered the question?

- What if the setting of this question was different, how does that impact your response?
    - *Further probe*: What rank would you have assigned each response if the question had been in a setting that was (non-)healthcare-related?
    - *Further probe*: Was there wording about this question that influenced your response?

- What knowledge or ability do you think this question is testing? Why do you think this?
    - *Alternative phrasing:* What do you think this question is asking you to do and why?

The interview may conclude with the following questions based on time:

- What questions do you feel were easiest to answer and why?
    - What questions do you feel were difficult to answer and why?
    - How did ranking each option influence your response?

- In general, what factors do you believe influenced your response to each scenario?

- If you had known all of these questions were testing empathy, how would that have changed your responses?

- How did knowing that this test may be used for residency selection influence your responses?

- What questions made you feel confused and why? Do you feel you did not understand some of the questions?

- How would your responses have differed if the questions were open-ended?

The last part of this session includes a brief 5-minute questionnaire. Once you have finished the questionnaire the session is complete. Thank you again for your participation.

APPENDIX L. FINAL CODEBOOK

Code (Abbv)	Description	Samples
Situational Judgment Test Framework		
Lack Experience (LE)*	Reference to not having witnessed or encountered a scenario or setting that is described.	"I can't think of a time…", "I have not seen this before…"
Nondescript Experience or Knowledge (NE)*	Memories, observations, facts, information, strategies, or skills provided without a clear distinction of the setting or environment in which they occurred.	"This has happened to me" (with no qualifiers to distinguish the setting), "This happens all the time"
General Experience (GE)	Memories or observations that are related to experiences outside of the health professions.	"I've had friends who have gone through loss", "Reminds me when I would vent to a friend"
General Knowledge (GK)	Facts, information, strategies, or skills identified to address problems that are encountered in contexts outside of the health professions and broadly applicable to societal or cultural expectations.	"I think there's social norms still… you're not going to let them start a fist fight in the grocery store", "We have university policies"
Specific Job Experience (JE)	Memories or observations that are related to experiences exclusively within the health professions.	"I remember a time in the hospital", "I work with patients who have depression every day"
Specific Job Knowledge (JK)	Facts, information, strategies, or skills identified to address problems that are encountered exclusively within the health professions.	"We are trained in mental health first aid", "We have a policy that", "It depends how they manage their diabetes"
Ability (AB)	Reference to the possession or lack of the means or skills to do something such as a talent, skill, or proficiency in a particular area.	"I don't know how to do that well", "I'm not really trained to…", "If I was more skilled at…"
Self-Perception (SP)	Awareness of the characteristics or qualities that form an individual's character or identity.	"As a pharmacist…", "It makes me uncomfortable", "I tend to be more judgy"
Emotional Intelligence (EI)	The capacity to be aware of, control, and express one's emotions and to handle relationships.	"They want you to validate their feelings", "That is upsetting"
Ability To Identify Construct (AC)	The examinee's attempt to identify which attribute is being evaluated by a test question.	"I think this is asking me to", "I'm not sure what I am expected to do here"
Impression Management (IM)	Extent to which an examinee modifies a response based on what is expected from the test administrator.	"The residency program director would want me to", "I'd want to look like I am compassionate"
Response Process Framework		
Comprehension (CO)	The cognitive process used by the examinee to read, interpret, or understand the purpose of the test item.	"The way I interpret this", "This sounds like", "I didn't read carefully"
Assumptions (AM)*	Interpretations or constraints placed on the scenario based on the perspective of the examinee.	"I assume this is said in a polite tone", "I think this comes of as…", "I am assuming there is…"
Objective (OB)*	Identification or prediction of a goal to be accomplished by a test item or response.	"What would be best for the patient", "The patient may take that in a bad way"
Retrieval (RT)	Accessing long-term memories and knowledge relevant to the scenario and proposed problem.	"This makes me think of…", "I remember…"
Judgment (JU)	Making a decision or value-statement; typically generated by integrating memories, knowledge, experiences, and other antecedents.	"This is a bad approach", "I think that is a good idea", "I would never do that", "Compared to this option", "You should...". NOT: "Yes", "No"
Feelings About the Test (FT)*	Emotions or comments regarding the quality of the test items.	"This one was difficult", "I didn't like…"
Perceptions (PR)*	Awareness of factors weighed when deciding the priority of response options.	"Often times they just want an apology", "How I would want to be treated in the situation"
Context (CT)*	Reference to how variations in the components of the scenario may affect the selected responses.	"It depends on…", "I don't think my answer would change in a healthcare setting"
Response Selection (RP)	The final verbal or written answer that is selected by the examinee.	"This would be number five", "It goes last"
Strategies (ST)*	Techniques used by examinees to answer test items.	"I selected the first and last option, then guessed"
Empathy Framework		
Affective Empathy (AE)	Individual's ability to experience and internalize the feelings experienced by others.	"They are likely upset or frustrated", "This is so sad"
Cognitive Empathy (CE)	Individual's ability to understand another person's perspective instead of being self-oriented.	"Trying to think about their perspective", "Putting myself in their shoes…"

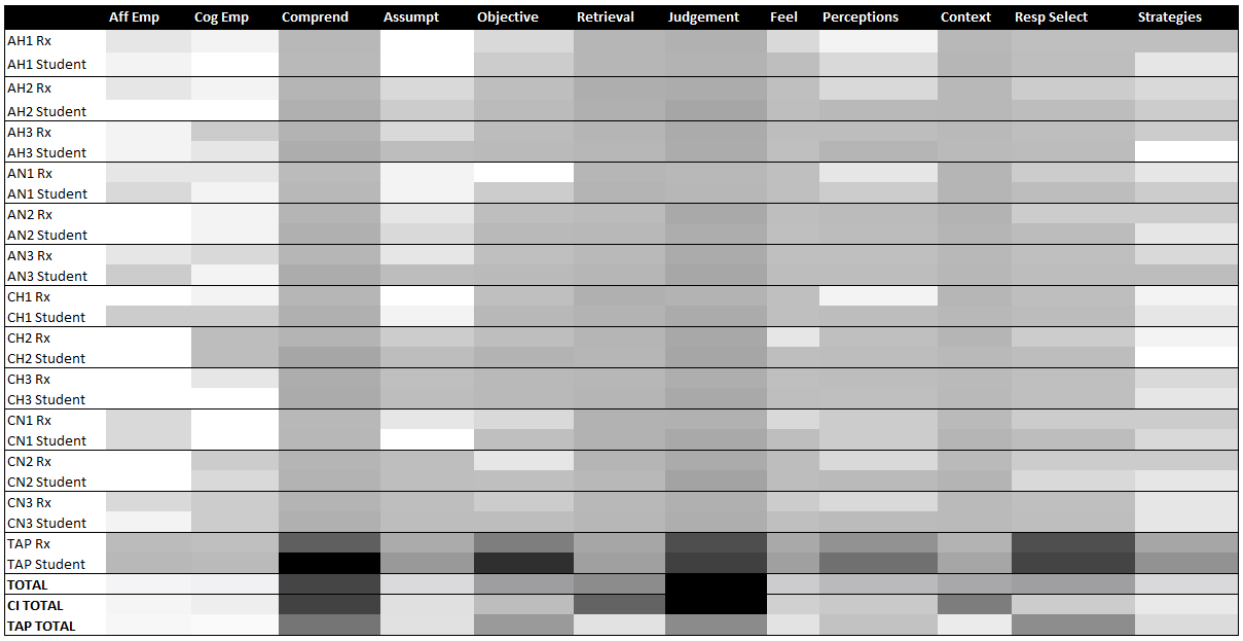
*****Code added through inductive process (i.e., not in the original codebook)

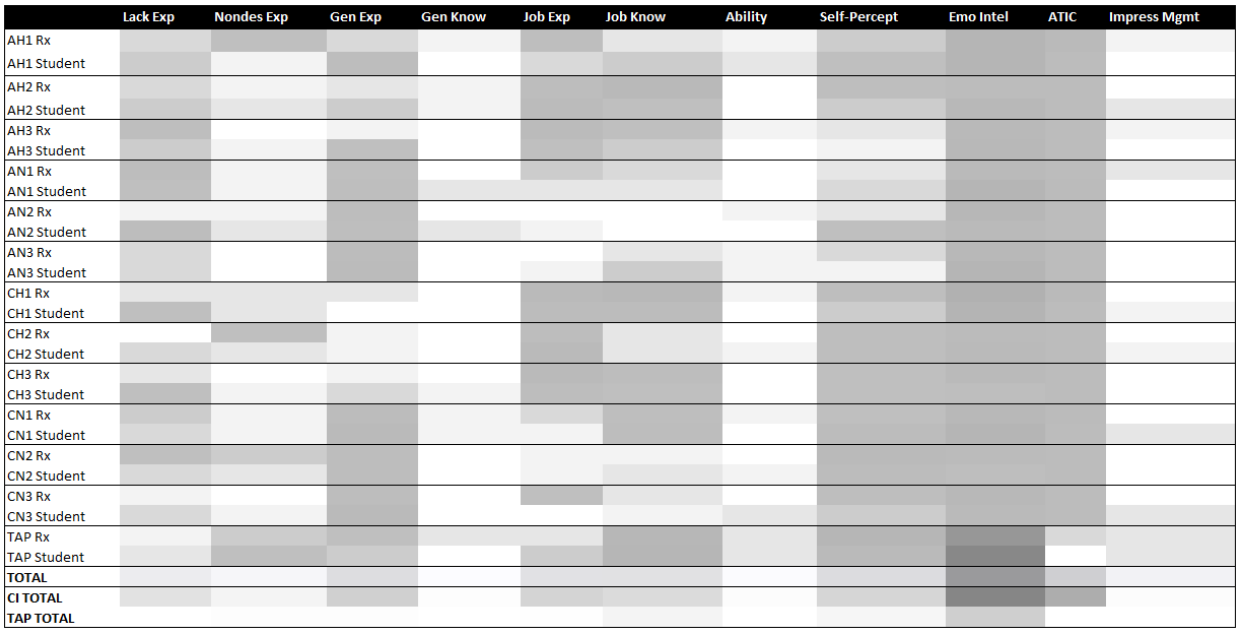
APPENDIX M. PARTICIPANT SJT PERFORMANCE DATA

Participant	CH1	CH2	CH3	CN1	CN2	CN3	AH1	AH2	AH3	AN1	AN2	AN3
P01	20	14	18	8	14	10	16	14	18	8	14	10
P02	16	18	18	10	12	14	12	20	18	10	18	8
P03	12	12	18	14	12	14	20	8	18	16	14	14
P04	12	16	18	12	16	10	16	14	14	18	20	18
P05	14	14	18	12	14	10	20	16	20	16	20	12
P06	20	20	18	12	14	10	20	8	18	16	18	16
P07	20	14	18	14	16	16	18	14	20	16	14	12
P08	20	14	14	16	10	14	14	12	16	12	16	16
P09	16	14	18	14	18	18	14	14	18	16	14	16
P10	16	12	18	20	16	12	18	12	18	12	16	12
P11	16	14	18	20	16	14	12	14	14	12	14	14
P12	12	14	12	12	18	16	14	8	16	18	16	12
P13	16	12	18	18	18	14	20	16	18	18	16	16
P14	14	14	12	14	16	20	20	12	20	10	14	20
P15	18	12	14	14	18	20	20	16	20	18	18	12
S01	12	12	20	14	12	10	12	16	18	18	14	16
S02	16	16	18	14	16	16	16	14	16	10	14	16
S03	18	20	12	14	18	16	8	12	18	16	14	14
S04	18	18	14	12	16	12	20	16	18	20	14	8
S05	12	20	14	14	14	10	14	14	16	10	16	16
S06	16	14	18	20	18	12	12	14	14	18	20	16
S07	12	16	18	14	16	20	12	12	18	16	18	16
S08	12	12	14	14	14	14	12	18	16	16	20	14
S09	10	12	18	14	10	8	12	12	14	8	10	14
S10	12	14	20	14	18	16	16	14	14	12	16	14
S11	16	20	18	10	10	14	12	12	18	18	12	14
S12	18	16	20	14	16	12	20	14	18	16	14	10
S13	20	14	18	12	14	12	12	8	18	14	14	16
S14	14	18	20	12	18	18	20	14	18	16	16	12
S15	20	16	20	12	12	16	18	8	18	18	12	8

APPENDIX N. HEAT MAP OF CODE DISTRIBUTION

Note: Shades of gray are indicative of the relative frequency of codes across the entire collection of interviews. White space indicates a low frequency and dark gray indicates a high frequency for colors pertaining to individual items.





REFERENCES

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Bardes, C. L., Best, P. C., Kremer, S. J., & Dienstag, J. L. (2009). Perspective: medical school admissions and non-cognitive testing: some open questions. *Academic Medicine, 84,* 1360-1363.

Bauer, T. N. & Truxillo, D. M. (2006). Applicant reactions to situational judgment tests: Research and related practical issues. In J. A. Weekly & R. E. Ployhart (Eds.), *Situational judgement tests: Theory, measurement, and application.* (1st ed., pp. 233-249), Mahwah, NJ: Lawrence Erlbaum.

Beier, M. E., & Ackerman, P. L. (2005). Age, ability, and the role of prior knowledge on the acquisition of new domain knowledge: Promising results in a real-world learning environment. *Psychology and Aging, 20,* 341-355.

Bergman, M. E., Drasgow, F., Donovan, M. A., Henning, J. B., & Juraska, S. (2006). Scoring situational judgment tests: Once you get the data, your troubles begin. *International Journal of Selection and Assessment, 14,* 223-235.

Berwick, D. M., & Finkelstein, J. A. (2010). Preparing medical students for the continual improvement of health and health care: Abraham Flexner and the new "public interest". *Academic Medicine, 85,* S56-S65.

Bond, T. G. & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). New York, NY: Routledge.

Bourdage, J. S., Wiltshire, J., & Lee, K. Personality and workplace impression management: Correlates and implications. (2015). *Journal of Applied Psychology, 100*(2), 537-546.

Brooks, M. E. & Highhouse, S. (2006). Can good judgment be measured? In J. A. Weekly & R. E. Ployhart (Eds.), *Situational judgement tests: Theory, measurement, and application.* (1st ed., pp. 39-56), Mahwah, NJ: Lawrence Erlbaum.

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology, 3,* 296-322.

Caines, J., Bridglall, B. L., & Chatterji, M. (2014). Understanding validity and fairness issues in high-stakes individual testing situations. *Quality Assurance in Education, 22*(1), 5-18.

Campion, M. C., Ployhart, R. E., & MacKenzie Jr., W. I. (2014). The state of research on situational judgment tests: A content analysis and directions for future research. *Human Performance, 27,* 283-310.

Catano, V. M., Brochu, A., & Lamerson, C. D. (2012). Assessing the reliability of situational judgment tests used in high-stakes situations. *International Journal of Selection and Assessment, 20*(3), 333-346.

Care, E., Scoular, C., & Griffin, P. (2016). Assessment of collaborative problem solving in education environments. *Applied Measurement in Education, 29*(4), 250-264.

Carre, A., Stefaniak, N., D'Ambrosio, F., Bensalah, L., & Besche-Richard, C. (2013). The Basic Empathy Scale in Adults (BES-A): Factor structure of a revised form. *Psychological Assessment, 25*(3), 679-691.

Chan, D. & Schmitt, N. (2002). Situational judgment and job performance. *Human Performance, 15*, 233-254.

Cheng, J. W., Chiu, W. L., Chang, Y. Y., & Johnstone, S. (2014). *Journal of Psychology*, *148*(6), 621-640.

Chessa, A. G., & Holleman, B. C. (2007). Answering attitudinal questions: Modelling the response process underlying contrastive questions. *Applied Cognitive Psychology, 21*, 203-225.

Chi, M. T. H. (1997). Quantifying qualitative analyses of verbal data: A practical guide. *The Journal of Learning Sciences, 6*, 271-315.

Christian, M. S., Edwards, B. D., & Bradley, J. C. (2010). Situational judgement tests: Constructs assessed and a meta-analysis of their criterion-related validities. *Personnel Psychology, 63*, 83-117.

Cizek, G. J. (2015). Validity test score meaning and defending test score use: different aims, different methods. *Assessment in Education: Principles, Policy, & Practice, 23*(2), 212-225.

Clevenger, J., Pereira, G. M., Wiechmann, D., Schmitt, N., & Harvey, V. S. (2001). Incremental validity of situational judgment tests. *Journal of Applied Psychology, 86*, 410-417.

Colbert-Getz, J. M., Pippitt, K., & Chan, B. (2015). Developing a situational judgment test blueprint for assessing the non-cognitive skills of applicants at the University of Utah School of Medicine, the United States. *Journal of Educational Evaluation for Health Professions, 12*, 51-55.

Cooke, M., Irby, D. B., & O'Brien, B. C. (2010). *Educating physicians: A call for reform of medical school and residency.* San Francisco, CA: Jossey-Bass.

Cowart, K., Dell, K., Rodriguez-Snapp, N., & Petrelli, H.M. (2016). An examination of correlations between MMI scores and pharmacy school GPA. *American Journal of Pharmaceutical Education, 80*(6), Article 98.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York, NY: Harcourt Brace Jovanovich College Publishers.

Crook, A. E., Beier, M. E., Cox, C. B., Kell, H. J., Hanks, A. R., & Motowidlo, S. J. (2011). Measuring relationships between personality, knowledge, and performance using single-response situational judgment tests. *International Journal of Selection and Assessment, 19*(4), 363-373.

Davis-Becker, S., & Muckle, T. J. (2017). Test design: Laying out the roadmap. In S. Davis-Becker & C. W. Buckendahl (Eds.), *Testing in the professions: Credentialing policies and practice* (1st ed., pp. 41-63), New York, NY: Routledge.

Decety, J. & Jackson, P. I. (2004). The functional architecture of human empathy. *Behavioral and Cognitive Neuroscience Reviews, 3*(2), 71-100.

De Leng, W. E., Stegers-Jager, K. M., Husbands, A., Dowell, J. S., Born, M. P., & Themmen, A. P. N. (2017). Scoring methods of a situational judgment test: Influence on internal consistency reliability, adverse impact and correlation with personality? *Advances in Health Sciences Education, 22,* 243-265.

Epstein, R. M., & Hundert, E. M. (2002). Defining and assessing professional competence. *Journal of the American Medical Association, 287,* 226-235.

Ercikan, K. & Oliveri, M. E. (2016). In search of validity evidence in support of the interpretation and use of assessments of complex constructs: Discussion of research on assessing 21st century skills. *Applied Measurement in Education, 29*(4), 310-318.

Ercikan, K. & Pellegrino, J. W. (2017). Validation of score meaning using examinee response process for the next generation of assessments. In K. Ercikan & J. W. Pellegrino (Ed.), *Validation of score meaning for the next generation of assessments: The use of response processes* (pp. 1-8). New York, NY: Routledge.

Fan, J., Stuhlman, M., Chen, L., & Weng, Q. (2016). Both general domain knowledge and situation assessment are needed to better understand how SJTs work. *Industrial and Organizational Psychology, 31*(1), 43-47.

Farrington, C. A., Roderick, M., Allensworth, E., Nagaoka, J., Keyes, T. S., Johnson, D. W., & Beechum, N. O. (2012). *Teaching adolescents to become learners. The role of noncognitive factors in shaping school performance: A critical literature review*. Chicago: University of Chicago Consortium on Chicago School Research.

Ferguson, E., & Lievens F. (2017). Future directions in personality, occupational, and medical selection: Myths, misunderstandings, measurements, and suggestions. *Advances in Health Sciences Education, 22,* 387-399.

Fjortoft, N., Van Winkle, L. J., & Hojat, M. (2011). Measuring empathy in pharmacy students. *American Journal of Pharmaceutical Education, 75*, Article 109.

Geisinger, K. F. (2016). 21st Century Skills: What are they and how do we assess them? *Applied Measurement in Education, 29*(4), 245-249.

Gerdes, K. E., Segal, E. A., & Lietz, C. A. (2010). Conceptualising and measuring empathy. *British Journal of Social Work, 40*, 2326-2343.

Gessner, T. L. & Klimoski, R. J. (2006). Making sense of situations. In J. A. Weekly & R. E. Ployhart (Eds.), *Situational judgement tests: Theory, measurement, and application.* (1st ed., pp. 13-38), Mahwah, NJ: Lawrence Erlbaum.

Griffin, B. (2014). The ability to identify criteria: Its relationship with social understanding, preparation, and impression management in affecting predictor performance in a high-stakes selection context. *Human Performance, 27*, 147-164.

Goldstein, E. A., Maestas, R. R., Fryer-Edwards, K., Wenrich, M. D., Oelschlarger, A. M., Baernstein, A., & Kimball, H. R. (2006). Professionalism in medical education: An institutional challenge. *Academic Medicine, 81*(10), 871-876.

Goss, B. D., Ryan, A. T., Waring, J., Judd, T., Chiavaroli, N. G., O'Brien, R. C., Trumble, S. C., & McColl, G. J. (2017). Beyond selection: The use of situational judgement tests in the teaching and assessment of professionalism. *Academic Medicine*. doi:10.1097/ACM.000000000000591 [epub ahead of print]

Guenole, N. Chernyshenko, O. S., & Weekly, J. (2017). On designing construct driven situational judgment tests: Some preliminary recommendations. *International Journal of Testing, 17*(3), 234-252.

Hafferty, F. W., O'Donnell, J. F., & Baldwin Jr., D. C. (2015). *The hidden curriculum in health professional education.* Hanover, NH: Dartmouth College

Haladyna, T. M. (1990). Effects of empirical option weighting on estimating domain scores and making pass/fail decisions. *Applied Measurement in Education, 3*(3), 231-244.

Hambrick, D. Z. (2003). Why are some people more knowledgeable than others? A longitudinal study of knowledge acquisition. *Memory and Cognition, 31*, 902-917.

Harris, A. M., Siedor, L. E., Fan, Y., Listyg, B., & Carter, N. T. (2016). In defense of the situation: An interactionist explanation for performance on situational judgment tests. *Industrial and Organizational Psychology, 31*(1), 23-28.

Harvey, R. J. (2016). Scoring SJTs for traits and situational effectiveness. *Industrial and Organizational Psychology, 31*(1), 63-71.

Hays, R. (2013). Assessing professionalism. In K. Walsh (Ed.), *Oxford textbook of medical education* (pp. 500-512). Oxford, UK: Oxford University Press.

Hojat, M. (2007). *Empathy in patient care: Antecedents, development, measurement, and outcomes.* New York, NY: Springer.

Hojat, M., Erdmann, J. B., Gonnella, J. S. (2013). Personality assessments and outcomes in medical education and the practice of medicine: AMEE Guide No. 79. *Medical Teacher, 35*, e1267-e1301.

Hojat, M., & LaNoue, M. (2014). Exploration and confirmation of the latent variable structure of the Jefferson scale of empathy. *International Journal of Medical Education, 5*, 73-81.

Hojat, M., Mangione, S., Nasca, T. J., Cohen, M. J. M., Gonnella, J. S., Erdmann, J. B., Veloski, J., & Magee, M. (2001). The Jefferson Scale of Physician Empathy: Development and preliminary psychometric data. *Educational and Psychological Measurement, 61*(2), 349-365.

Irby, D. M. (2011). Educating physicians for the future: Carnegie's call for reform. *Medical Teacher, 33*(7), 547-550.

Jolliffe, D., & Farrington, D. P. (2006). Development and validation of the Basic Empathy Scale. *Journal of Adolescence, 29*, 589-611.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*(3), 527-535.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17-64). Westport, CT: American Council on Education / Praeger.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1-73.

Kane, M. (2016). Validation strategies: Delineating and validating proposed interpretations and uses of test scores. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development* (2nd ed., pp. 64-80), New York, NY: Routledge.

Kane, M. T., Clauser, B. E., & Kane, J. (2017). Validation framework for credentialing tests. In S. Davis-Becker & C. W. Buckendahl (Eds.), *Testing in the professions: Credentialing policies and practice* (1st ed., pp. 21-40), New York, NY: Routledge.

Kane, M. T. & Mislevy, R. (2017). Validating score interpretations based on response processes. In K. Ercikan & J. W. Pellegrino (Ed.), *Validation of score meaning for the next generation of assessments: The use of response processes* (pp. 11-24). New York, NY: Routledge.

Keppel, G. (1991). *Design and analysis: A researcher's handbook* (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Kim, S. S., Kaplowitz, S., & Johnston, M. V. (2004). The effects of physician empathy on patient satisfaction and compliance. *Evaluation and the Health Professions, 27*(3), 237-251.

Kleinmann, M., Ingold, V. P., Lievens, F., Jansen, A., Melchers, K. G., & Konig, C. J. (2011). A different look at why selection procedures work: The role of candidates' ability to identify criteria. *Organizational Psychology Review, 1*, 128-46.

Koczwara, A., Patterson, F., Zibarras, L., Kerrin, M., Irish, B., & Wilkinson, M. (2012). Evaluating cognitive ability, knowledge tests, and situational judgment tests for postgraduate selection. *Medical Education, 46*, 399-408.

Krumm, S., Lievens, F., Huffmeier, J. Lipnevich, A. A., Bendels, H., & Hertel, G. (2015). How "situational" is judgment in situational judgment tests? *Journal of Applied Psychology, 100*, 399-416.

Lane, S., Raymond, M. R., Haladyna, T. M., & Downing, S. M. (2016). Test development process. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development* (2nd ed., pp. 3-18), New York, NY: Routledge.

Larson, E. B. & Yao, X. (2005). Clinical empathy as emotional labor in the patient-physician relationship. *Journal of the American Medical Association, 293*, 1100-1106.

Leighton, J. P. (2017a). *Using think-aloud interviews and cognitive labs in educational research: Understanding qualitative research.* New York, NY: Oxford University Press.

Leighton, J. P. (2017b). Collecting and analyzing verbal response process data in the service of interpretative and validity arguments. In K. Ercikan & J. W. Pellegrino (Ed.), *Validation of score meaning for the next generation of assessments: The use of response processes* (pp. 25-38). New York, NY: Routledge.

Leighton, J. P. & Gierl, M. J. (2011). *The learning sciences in educational assessment: The role of cognitive models*. New York, NY: Cambridge University Press.

Levine, R. B., & Cayea, D. (2015). Defining and assessing the 21st-century physician in training. *Journal of General Internal Medicine, 30*, 1241-1242.

Li, H., Ding, N., Zhang, Y., Liu, Y., & Wen, D. (2017). Assessing medical professionalism: A systematic review of instruments and their measurement properties. *PLoS ONE, 12*(5), e0177321. Doi:10.1371/journal.pone.0177321.

Lievens, F. (2017). Construct-driven SJTs: Toward and agenda for future research. *International Journal of testing, 17*(3), 269-276.

Lievens, F., Buyse, T., Sackett, P. R., & Connelly, B. S. (2012). The effects of coaching on situational judgment tests in high-stakes selection. *International Journal of Selection and Assessment, 20*(3), 272-282.

Lievens F., & Motowidlo, S. J. (2016). Situational judgment tests: From measures of situational judgment to measures of general domain knowledge. *Industrial and Organizational Psychology, 9*(1), 3-22.

Lievens, F., & Patterson, F. (2011). The validity and incremental validity of knowledge tests, low-fidelity simulations, and high-fidelity simulations for predicting job performance in advanced-level high-stakes selection. *Journal of Applied Psychology,* 96, 927-940.

Lievens, F., Peeters, H., & Schollaert, E. (2008). Situational judgment tests: A review of recent research. *Personnel Review, 37*(4), 426-441.

Lievens, F. & Peeters, H. (2009). Impact of elaboration on responding to situational judgment test items. *International Journal of Selection and Assessment, 16*, 345-355.

Luecht, R. M. (2017). Data and scale analysis for credentialing examinations. In S. Davis-Becker & C. W. Buckendahl (Eds.), *Testing in the professions: Credentialing policies and practice* (1st ed., pp. 123-152), New York, NY: Routledge.

Marentette, B. J., Meyers, L. S., Hurtz, G. M., & Kuang, D. C. (2012). Order effects on situational judgment tests items: A case of construct-irrelevant difficulty. *International Journal of Selection and Assessment, 20*(3), 319-332.

McDaniel, M. A., Hartman, N. S., Whetzel, D., & Grubb III, W. L. (2007). Situational judgment tests, response instructions, and validity: A meta-analysis. *Personnel Psychology, 60*(1), 63-91.

McDaniel, M. A., List, S. K., & Kepes, S. (2016). The "hot mess" of situational judgment test construct validity and other issues. *Industrial and Organizational Psychology, 31*(1), 47-51.

McDaniel, M. A., Morgenson, F. P., Finnegan, E. B., Campion, M. A., & Braverman, E. P. (2001). Use of situational judgment tests to predict job performance: A clarification of the literature. *Journal of Applied Psychology, 80,* 730-740.

McDaniel, M. A., & Nguyen, N. T. (2001). Situational judgment tests: A review of practice and constructs assessed. *International Journal of Selection and Assessment, 9*(1/2), 103-113.

McDaniel, M. A., Whetzel, D. L., Hartman, N. S., Nguyen, N. T., & Grubb, W. L. (2006). Situational judgment tests: Validity and an integrative model. In J. A. Weekly & R. E. Ployhart (Eds.), *Situational judgement tests: Theory, measurement, and application.* (1st ed., pp. 183-203), Mahwah, NJ: Lawrence Erlbaum.

Melchers, K. G., & Kleinmann, M. (2016). Why situational judgment is a missing component in the theory of SJTs. *Industrial and Organizational Psychology, 31*(1), 29-34.

Merriam, S. B. & Tisdell, E. J. (2016). *Qualitative research: A guide to design and implementation.* (4th ed.). San Francisco: Jossey-Bass.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-104), New York: American Council on Education / Macmillan.

Miller, G. E. (1990). The assessment clinical skills/competence/performance. *Academic Medicine, 65*(9 Suppl), S63-S67.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). A brief introduction to evidence-centered design. *ETS Research Report Series, 1*, 1-29.

Morse, J. M., Anderson, G., Bottorff, J. L., Yonge, O., O'Brien, B., Solberg, S. M., & McIlveen, K. H. (1992). Exploring empathy: A conceptual fit for nursing practice? *Image: The Journal of Nursing Scholarship, 24*(4), 273-280.

Motowidlo, S. J. & Beier, M. B. (2010). Differentiating specific job knowledge from implicit trait policies in procedural knowledge measured by a situational judgment test. *Journal of Applied Psychology, 95*(2), 321-333.

Motowidlo, S. J., Crook, A. E., Kell, H. J., & Naemi, B. (2009). Measuring procedural knowledge more simply with a single-response situational judgment test. *Journal of Business Psychology, 24*(3), 281-288.

Motowidlo, S. J., Dunnette, M. D., & Carter, G. W. (1990). An alternative selection procedure: The low-fidelity simulation. *Journal of Applied Psychology, 75*(6), 640-647.

Motowidlo, S. J., Ghosh, K., Mendoza, A. M., Buchanan, A. E., & Lerma, M. N. (2016). A context-independent situational judgment test to measure prosocial implicit trait policy. *Human Performance, 29*(4), 331-346.

Motowidlo, S. J., Hooper, A. C., & Jackson, H. L. (2006). Implicit policies about relations between personality traits and behavioral effectiveness in situational judgment items. *Journal of Applied Psychology, 91*(4), 749-761.

National Research Council. (1999). *How people learn: Bridging research and practice.* Committee on Learning Research and Educational Practice. M. S. Donovan, J. D. Bransford, and J. W. Pellegrino, (Eds.). Commission on Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.

Neufeld, V. R., & Norman, G. R. (1985). *Assessing Clinical Competence.* New York: Springer-Verlag.

Newell, A. & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

Nguyen, N. T., Biderman, M. D., & McDaniel, M. A. (2005). Effects of response instructions on faking a situational judgment test. *International Journal of Selection and Assessment, 13*, 250-260.

Nichols, P. & Huff, K. (2017). Assessments of complex thinking. In K. Ercikan & J. W. Pellegrino (Ed.), *Validation of score meaning for the next generation of assessments: The use of response processes* (pp. 63-74). New York, NY: Routledge.

Nunes, P., Williams, S., Sa. B., & Stevenson, K. (2011). A study of empathy decline in students from five health disciplines during their first year of training. *International Journal of Medical Education, 2*, 12-17.

Oranje, A., Gorin, J., Jia, Y., & Kerr, D. (2017). Collecting, analyzing, and interpreting response time, eye-tracking, and log data. In K. Ercikan & J. W. Pellegrino (Ed.), *Validation of score meaning for the next generation of assessments: The use of response processes* (pp. 39-51). New York, NY: Routledge.

Patterson, F., Ashworth, V., Kerrin, M., O'Neill, P. (2013). Situational judgment tests represent a measurement method and can be designed to minimize coaching effects. *Medical Education, 47*(2), 220-221.

Patterson, F., Ashworth, V., Zibarras, L., Coan, P., Kerrin, M., & O'Neill, P. (2012) Evaluations of situational judgement tests to assess non-academic attributes in selection. *Medical Education, 46*, 850-868.

Patterson, F., Baron, H., Carr, V., Plint, S., & Lane, P. (2009). Evaluation of three short-listing methodologies for selection into post-graduate training: The case of General Practice in the UK. *Medical Education, 43*, 50-57.

Patterson, F., Cleland, J., & Cousans, F. (2017). Selection methods in healthcare professions: where are we now and where next? *Advances in Health Sciences Education, 22*(2), 229-242.

Patterson, F., Knight, A., Dowell, J., Nicholson, S., Cousans, F., & Cleland, J. (2016). How effective are selection methods in medical education? A systematic review. *Medical Education, 50*, 36-60.

Patterson, F., Zibarras, L., & Ashworth, V. (2016). Situational judgement tests in medical education and training: Research, theory, and practice: AMEE Guide No. 100. *Medical Teacher, 38*(1), 3-17.

Pellegrino, J., Chudowsky, N., & Glaser, R. (Committee on the Foundations of Assessment) (Eds.) (2001). *Knowing what students know: The science and design of educational assessment.* Board on Testing and Assessment, Center for Education, National Research Council, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academic Press.

Petty-Saphon, K., Walker, K. A., Patterson, F., Ashworth, V., & Edwards, H. (2016). Situational judgment tests reliably measure professional attributes important for clinical practice. *Advances in Medical Education and Practice, 8,* 21-23.

Persky, A. M., Greene, J. M., Anksorus, H., Fuller, K. A., & McLaughlin, J. E. (2018). Developing an innovative, comprehensive first-year capstone to assess and inform student learning and curriculum effectiveness. *American Journal of Pharmaceutical Education. [epub ahead of print]*

Ployhart, R. E. (2006). The predictor response process model. In J. A. Weekly & R. E. Ployhart (Eds.), *Situational judgement tests: Theory, measurement, and application.* (1st ed., pp. 83-105), Mahwah, NJ: Lawrence Erlbaum.

Ployhart, R. E., & Erhart, M. G. (2003). Be careful what you ask for: Effects of response instructions on the construct validity and reliability of situational judgment tests. *International Journal of Selection and Assessment, 11,* 1-16.

Quince, T., Thiemann, P., Benson, J., & Hyde, S. (2016). Undergraduate medical students' empathy: Current perspectives. *Advances in Medical Education and Practice, 7,* 443-455.

Ratanawongsa, N., Bolen, S., Howell, E. E., Kern, D. E., Sisson, S. D., & Larriviere, D. (2006). Residents' perceptions of professionalism in training and practice: Barriers, promoters, and duty hour requirements. *Journal of General Internal Medicine, 21*(7), 758-763.

Rees, E. L., Hawarden, A. W., Dent, G., Hays, R., Bates, J., & Hassell, A. B. (2016). Evidence regarding the utility of multiple mini-interview (MMI) for selection to undergraduate health programs: A BEME systematic review. BEME Guide No. 37. *Medical Teacher, 38*(5), 443-455.

Renate, L. E. P. R., Corcoran, R., Drake, R., Shryane, N. M., & Vollm, B. A. (2011). The QCAE: A questionnaire of cognitive and affective empathy. *Journal of Personality Assessment, 93*(1), 84-95.

Riconscente, M. M., Misley, R. J., & Corrigan, S. (2016). Evidence-centered design. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development* (2nd ed., pp. 40-63), New York, NY: Routledge.

Rockstuhl, T., Ang S., Ng, K. Y., Lievens, F., and Van Dyne, L. (2015). Putting judging situations into situational judgment tests: Evidence from intercultural multimedia SJTs. *Journal of Applied Psychology, 100,* 464-480.

Rogers, C. (1951). *Client-centered therapy.* London, England: Constable.

Saldana, J. (2016). *The coding manual for qualitative researchers* (3rd ed.). Thousand Oaks, CA: SAGE.

Schwarz, N. (2007). Cognitive aspects of survey methodology. *Applied Cognitive Psychology, 21*, 277-287.

Siebert, C. F., & Siebert, D. C. (2018). *Data analysis with small samples and non-normal data* (1st ed.), New York, NY: Oxford University Press.

Siegel, S., & Castellan, N. J., Jr. (1988). *Nonparametric statistics for the behavioral sciences* (2nd ed.), New York, NY: McGraw-Hill.

Sorrel, M. A., Olea, J., Abad, F. J., de la Torre, J., Aguado, D., & Lievens, F. (2016). Validity and reliability of situational judgment test scores: A new approach based on cognitive diagnosis models. *Organizational Research Methods, 19*(3), 506-532.

Spearman, C. (1910). Correlation calculated with faulty data. *British Journal of Psychology, 3*, 271-295.

St-Sauveur, C., Girouard, S. & Goyette, V. (2014). Use of situational judgment tests in personnel selection: Are the different methods for scoring the response options equivalent? *International Journal of Selection and Assessment, 22*(3) ,225-239.

Tamayo, C. A., Rizkalla, M. N., & Henderson, K. K. (2015). Cognitive, behavioral, and emotional empathy in pharmacy students: Targeting programs for curriculum modification. *Frontiers in Pharmacology, 7,* Article 96.

Tourangeau, R., Rips, L. C., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge, MA: Cambridge University Press.

Truxillo, D. M., Bauer, T. N., Campion, M. A., & Paronto, M. E. (2002). Selection fairness information and applicant reactions: A longitudinal field study. *Journal of Applied Psychology, 87,* 1020-1031.

Weekley, J. A., & Ployhart, R. E. (2005). Situational judgment: Antecedents and relationships with performance. *Human Performance, 18,* 81-104.

Weekley, J. A., & Ployhart, R. E. (2006). An introduction to situational judgment testing. In J. A. Weekly & R. E. Ployhart (Eds.), *Situational judgement tests: Theory, measurement, and application.* (1st ed., pp. 1-10), Mahwah, NJ: Lawrence Erlbaum.

Weekley, J., Ployhart, R. E., & Holtz, B. C. (2006). On the development of situational judgment tests: Issues in item development, scaling, and scoring. In J. A. Weekly & R. E. Ployhart

(Eds.), *Situational judgement tests: Theory, measurement, and application.* (1st ed., pp. 157-182), Mahwah, NJ: Lawrence Erlbaum.

Whetzel, D. L., & McDaniel, M. A. (2009). Situational judgment tests: An overview of current research. *Human Resources Management Review, 19,* 188-202.

Willis, G. B. (2015). *Analysis of the cognitive interview in questionnaire design: Understanding qualitative research.* New York, NY: Oxford University Press.

Wolcott, M. W., Lupton-Smith, C., Cox, W. C., & McLaughlin, J. E. (2018). A 5-minute situational judgment test to assess empathy in first year student pharmacists. *American Journal of Pharmaceutical Education*. doi.org/10.5688/ajpe6960.

Wu, B., Wang, M., Grotzer, T. A., Liu, J., Johnson, J. M. (2016). Visualizing complex processes using a cognitive-mapping tool to support the learning of clinical reasoning. *BMC Medical Education, 16*(1), 216.