Understanding Faculty Assessment Decisions of

Medical Student Clinical Reasoning Ability

By

MARJORIE WESTERVELT
DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Education

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Lee Martin, Chair

Megan Welsh

Samuel Clarke

Committee in Charge

2021

Acknowledgments

I am forever grateful for the support I have received from my colleagues and mentors in the UC Davis School of Medicine. Without their guidance and encouragement, pursuing this degree would not have been possible. Thank you for always challenging me to grow.

Thank you to the faculty in the School of Education who have inspired me to think differently about medical education and pursue a career of educational research. I would not be here without the constant guidance from my advisor, Lee Martin, who took a chance in accepting me and my unique research interest. I am grateful for your willingness to discuss all my crazy ideas and for pushing me to constantly reassess my thinking. Thank you to Megan Welsh for your generous encouragement and for inspiring my passion for assessment and measurement. Thank you to Sam Clarke, for agreeing to join me through this program and providing an invaluable bridge between my work in education and medicine.

There are not enough words to express the depth of my love and gratitude for my husband, Michael. Thank you for sacrificing countless weekends and evenings to allow me to pursue this degree while working full time. Your love and support means everything to me, and I would not have been able to get through it without you.

Table of Contents

Abstract

Clinical reasoning is a fundamental skill required of all physicians. Direct observation is one method medical schools use to assess clinical reasoning, where faculty observers rate students based on the student's interaction with a patient. Variability in how individual faculty members define clinical reasoning, however, can reduce assessment reliability. Understanding how faculty make assessment decisions of student clinical reasoning can improve the reliability and validity of medical school's assessments. Fourteen UC Davis School of Medicine faculty members completed think-aloud interviews while watching a medical student encounter with a standardized patient. Faculty members were asked to assess the student's clinical reasoning ability and were not provided any information about the student or the case other than a door note. The faculty were then asked to provide written summative feedback to the student. The think-aloud interviews were video-recorded, transcribed, and analyzed using thematic analysis. The analysis provided five themes about how faculty members assess medical students: student factors, situational factors, assessor factors, integration, and judgment. Additional findings about the ways in which faculty provide students narrative feedback were also noted. The themes together create a model of faculty reasoning, the process by which faculty make assessment decisions about a medical student's clinical reasoning ability. Faculty assessment decisions are influenced by a number of different factors. The ways in which they process information about the student and the encounter, and then integrate it with their own existing knowledge and experience, is unique to the individual. Understanding this process allows for opportunities to influence factors to improve consistency, and therefore validity.

Understanding Faculty Assessment Decisions of Medical Student Clinical Reasoning Ability

I. Introduction

To take care of patients effectively physicians must understand vast amounts of biomedical knowledge and be able to apply this knowledge in front of patients in the clinical setting. In 2000, the Institute of Medicine (IOM) published a report highlighting the increase in preventable medical errors over the last several decades, with error in diagnosis accounting for 17% of those errors among practicing physicians outside of training (Donaldson et al., 2000). Twenty years later, a healthcare safety nonprofit, ECRI, rated “missed and delayed diagnoses” as the top concern for patient safety (ECRI, 2020). So how do physicians acquire the skills and knowledge they need to make correct diagnoses? And what assessment systems are in place to ensure they have these skills before they enter practice? I argue that improving and calibrating current medical student assessment practices will have lasting implications on medical students as they continue in their training. We can only improve these assessment practices, however, when we truly understand how assessments work and how assessment decisions are made.

To become a doctor, medical students must complete a grueling curriculum that challenges them to think differently and to adopt the behaviors of medical professionals. The three to four years that make up medical school, or undergraduate medical education (UME), are filled with continuous assessment, national licensing examinations, and immersive clinical experiences. Upon graduation from medical school, students then enter a second, more advanced phase of training called residency, where they gain more autonomy and are able to see patients with indirect supervision. To ensure medical students are ready to make this transition, schools rely on numerous assessments that are meant to determine whether medical students can take the knowledge they are acquiring and make decisions in real-world patient encounters.

In practice, medical decision-making is most often done at the bedside or in the exam room (Higgs et al., 2018). In almost every clinical encounter, physicians gather information about their patients by reviewing their medical records, collecting a history, and performing a physical exam. They will use this information to form a diagnosis, which in turn will lead them to create a plan of treatment or request additional testing (e.g. imaging, lab work, etc.) to confirm what they think may be happening. The process of deciding upon a diagnosis and implementing a treatment strategy is known as clinical reasoning (Durning et al., 2013).

While clinical reasoning has been of interest to medical education researchers for over 30 years, the medical education community still struggles to accurately assess the clinical reasoning abilities of its medical trainees (Norman, 2005). Various forms of assessment have been developed and implemented in medical schools, including, but not limited to, multiple choice question (MCQ) exams, objective structured clinical encounters (OSCEs), reviews of clinical notes, and performance assessments on clinical rotations. However, the reliability, or internal consistency (Bandalos, 2018), and validity, or "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (Bandalos, 2018, p. 255), of these stand-alone assessments has come into question when considering the multi-faceted nature of the clinical reasoning process (Daniel et al., 2019; Durning et al., 2013). Without an accurate assessment of whether or not medical students, also referred to as "learners" or "trainees", are developing clinical reasoning skills, UME programs risk advancing students to residency without a solid foundation for medical-decision making. An assessment's accuracy, however, is dependent on the individual using the instrument as well as the instrument itself.

Past research has identified a number of factors that influence how faculty make assessment decisions based on their observations of residents (more advanced learners) (Gingerich et al., 2014), but this research was not on the assessment of clinical reasoning. Furthermore, the physicians who are assessing medical students, both within academic institutions and in community settings, do not have a shared mental model of what clinical reasoning is, nor do they share an agreed upon framework for defining and assessing what clinical reasoning is (Durning et al., 2013). This raises worrisome arguments about the reliability and validity of the assessments being completed, as well as issues around fairness if students who perform equally are being evaluated against different standards. The inability for a faculty member to define what they are assessing when they observe a student also makes it extremely difficult for them to provide good, actionable feedback that "fosters ongoing learning" (Norcini et al., 2011).

To improve the assessment of clinical reasoning, we must first understand how faculty are making assessment decisions about students. The following study attempted to answer the following research questions:

1) How are faculty making assessment decisions about medical students' clinical reasoning abilities?
2) What commonalities or differences exist between different faculty members' reasoning processes and what opportunities exist to influence those differences?

To answer these research questions and understand faculty reasoning I used think-aloud interviews of faculty observing a video-recorded encounter of students in an objective structured clinical examination (OSCE). Faculty were also asked to provide written summative comments of the student's performance. I analyzed both the think-aloud interview transcripts and the summative comments using thematic analysis, an inductive qualitative research method. The results of the analysis identified five themes, which together create a framework for understanding faculty reasoning. Lastly, I explore opportunities to influence these themes to improve the reliability and validity of clinical reasoning assessment, as well as potential areas for future research.

II. Review of Literature

The following sections will explore what clinical reasoning is and the various methods in which medical educators and researchers have tried to assess it. In addition, I will give an overview of think-aloud interviews and thematic analysis, the two primary methods used to undertake this study.

What is Clinical Reasoning

Higgs, Jensen, Loftus, and Christensen (2018) argue that "clinical reasoning is the foundation of professional clinical practice" (p. xiii). Despite its importance within the medical field, as well as other health professions, a single model or definition of what clinical reasoning is, has not yet been agreed upon (M. Young et al., 2018). In a scoping review of clinical reasoning assessments, Daniel et al. (2019) defined clinical reasoning as "a skill, process, or outcome, wherein clinicians observe, collect, and interpret data to diagnose and treat patients" (p. 7). The authors broke apart clinical reasoning into seven distinct components: information gathering; hypothesis generation; problem representation; differential diagnosis; leading or working diagnosis; diagnostic justification; and management and treatment (Daniel et al., 2019). These components are consistent with other definitions and representations of clinical reasoning presented within the medical education literature (Daniel et al., 2019; Higgs et al., 2018; Norman, 2005).

As outlined above, information gathering is the first step in the clinical reasoning process (Daniel et al., 2019). Information gathering includes the collection of data from a variety of different sources; two sources include the patient and the medical record. During a clinical encounter, a clinician gathers information from the patient by taking a history, performing a physical exam, and ordering diagnostic tests such as bloodwork or imaging. Physicians also access the medical record to identify any past diagnoses or medical history that could still be affecting the patient’s health. Once information is collected, students must organize this data so it can be used to determine a diagnosis and treatment strategy. To do so, new information must be compared with the biomedical knowledge they have already acquired. This brings students into the hypothesis generation phase (Daniel et al., 2019).

Clinical Reasoning Strategies

There are three foundational clinical reasoning strategies within the medical education literature that move a physician or medical student through the hypothesis generation and diagnosis phases: hypothetico-deductive reasoning, scheme-inductive reasoning, and pattern recognition (Coderre et al., 2003). The hypothetico-deductive process is not unique to medicine, as it is the core of the scientific method in which a hypothesis is generated and data is collected to confirm or reject the hypothesis (Elstein, 1994). However, it became a leading theory of how physicians think through diagnosing patients in the early-1970s with the publication of the Medical Inquiry Project (Elstein et al., 1972). This study determined that based on the initial information collected from their patients, physicians generate several diagnoses and use these to guide what data they collect to confirm or reject those diagnoses (Elstein, 1994). Scheme-inductive reasoning, unlike hypothetico-deductive reasoning, begins with a single diagnosis rather than multiple (Coderre et al., 2003). That diagnosis is attached to a “scheme” or knowledge map of the different symptoms and probabilities. Lastly, physicians use pattern recognition when a patient’s symptoms and presentation are similar to a patient they have seen in the past, and they can therefore make a diagnosis without further testing (Higgs et al., 2018). Both hypothetico-deductive and scheme-inductive reasoning are referred to as analytical forms of clinical reasoning whereas pattern recognition is referred to as non-analytical (Norman et al., 2007).

All three of these processes are complex in that they require physicians to draw on a deep fund of knowledge and experience, but at their core, these clinical reasoning strategies employ the same cognitive processes used in critical thinking. Critical thinking has been defined as "purposeful, self-regulatory judgement which results in interpretation, analysis, evaluation, and inference, as well as explanation of the evidential, conceptual, methodological, criteriological, or contextual considerations upon which that judgement is based" (Facione, 1990, p. 3). For a single individual to think critically, he or she must engage in multiple complex thought or cognitive processes, most notably evaluation, analysis, and inference (Dwyer et al., 2014). To better understand how one engages in critical thinking, and ultimately clinical reasoning, I will take a deeper dive into three underlying cognitive processes: metacognition, deductive reasoning, and analogical reasoning.

Metacognition is the process of an individual introspectively looking at his or her own thoughts or beliefs, and in doing so adapting or adjusting his or her behaviors or decisions. Fletcher and Carruthers (2012) explored the cognitive structure that enables self-reflection and identified two systems, system 1 and system 2. System 1 provides swift, immediate judgments with little to no effort. System 2, on the other hand is slower and more mindful (Fletcher & Carruthers, 2012). Because it requires conscious thought, the authors argued that system 2 can be used to regulate reactions provoked by system 1. In comparing the definition of critical thinking above to the description of system 2 provided by Fletcher and Carruthers (2012), metacognition is vital for critical thinking because it allows individuals to reflect on what is being presented to them and to question their own thoughts and beliefs. For example, while a physician may start to enable pattern recognition to make a decision using their intuition (system 1), they can stop that instinct and reconsider the task at hand more thoughtfully with a hypothetico-deductive approach (system 2).

This thoughtful, system 2 decision-making may call upon deductive reasoning, another process underlying clinical reasoning. Deductive reasoning allows individuals to build conclusions based on true statements (Johnson-Laird, 1999). This ground-up approach means that if any underlying assumption is not true, then any assertions that build off of that assumption must also be false. Johnson-Laird (1999) argued that the strongest theory to explain deductive reasoning uses mental models. A mental model is a psychological construction of how different things, concepts, or components of a system interact or relate to one another (Collins & Gentner, 1987). As individuals critique an argument or assertion, they are likely to compare it with the mental models that they have already formed to determine if it is true or false, similar to physicians or medical students comparing new information or data about a patient to a proposed diagnosis. Proponents of the mental model theory argue that these models can be influenced by knowledge or biases, which explains how people presented with the same problem can come up with different solutions (Johnson-Laird, 1999). Using this argument, one can see how medical students employing hypothetico-deductive reasoning may come to different a conclusion than practicing physicians or even more experienced learners. As their knowledge and experience grows, their mental models change and they interpret new information differently when determining a final diagnosis.

Analogical reasoning is similar to deductive reasoning in that it also uses mental models, for individuals to make inferences about things with which they are unfamiliar (Collins & Gentner, 1987). According to Gentner (1983), structure-mapping theory claims that the use of analogy is grounded in relations between objects and that the way these relationships are mapped are dependent on higher-order, systematic relations. Individuals use analogy to apply already formed relationships between items to new items that are similar (Collins & Gentner, 1987). These relational structures, or mental models, allow them to then make decisions. For physicians using scheme-inductive reasoning, these mental models of existing knowledge of disease are compared with new findings and symptoms to infer the proper diagnosis and treatment.

Metacognition, deductive reasoning and analogical reasoning work in conjunction when an individual undertakes critical thinking, and therefore also clinical reasoning. When examining them together, they offer explanations of how individuals interpret information and make decisions or conclusions, even if the content is new or unfamiliar. For example, when an individual is presented with a new argument, deductive reasoning allows him or her to examine, or analyze, its foundations to assess credibility. Analogical reasoning then allows the individual to call upon previously constructed mental models to make inferences. Lastly, individuals can evaluate the validity of those inferences through metacognition. A similar process can be mapped to a physician determining a diagnosis: when presented with a new patient presentation, they can draw upon already existing medical knowledge to analyze the problem or symptom, use the relations of that knowledge with new information they gather to make inferences about a correct diagnosis, and then check their reasoning to ensure they've arrived at the appropriate conclusion.

As demonstrated above, the reasoning processes that allow an individual to successfully engage in critical thinking also affect a physician's ability to make medical decisions. Following a very deductive approach, hypothetico-deductive and scheme-inductive reasoning rely on slow and deliberate system 2 thinking. Pattern recognition uses the resolution of past problems to inform addressing new, but similar problems, which is more similar to system 1 thinking where decisions are made based on intuition. Proponents of dual-process theory argue that physicians engage in both analytical and non-analytical reasoning during the clinical reasoning process (Surry et al., 2017).

Influence of experience on clinical reasoning ability

Coderre et al. (2003) found that pattern recognition had the greatest likelihood of diagnostic success. One caveat to these findings, however, is that an individual can only utilize the pattern recognition strategy once they have been exposed to enough examples of cases with similar features that patterns can emerge. This is not possible for medical students who have yet to see patients on their own. In exploring the effectiveness of non-analytical reasoning between novices (i.e. medical students) and experts (i.e. practicing clinicians), Norman, Young, and Brooks (2007) found that, as expected, expert dermatologists have more diagnostic success than novices. They argued that this is because experts have accumulated more experience and seen more cases, which has allowed them to form relationships between certain attributes and diagnoses. They defined expertise as the "process of gradually acquiring more and more complex and better and better calibrated rules relating a set of characteristic attributes (signs and symptoms) to categories (diagnoses)" (Norman et al., 2007, p. 174). It was only on atypical cases where experts failed. Novices had difficulty on the same issues, demonstrating a lack of pattern recognition for rare or unusual presentations that led to the misdiagnosis, demonstrating that if a medical student were to try to use pattern recognition as a strategy for medical decision-making, but did not yet have enough cases to inform that decision-making process, they would likely come to the wrong conclusion. One approach to expertise development is deliberate practice, which is practice with intentional feedback and repeated opportunities to practice and improve. Over time, deliberate practices has been shown to be a promising method for enhancing the development of pattern recognition in medical students and professionals (K. A. Ericsson, 2004).

The fact that experts and novices perform differently on clinical reasoning tasks suggests that clinical reasoning is a skill that can be learned and developed over time. Guerrasio and Lessing (2018) argued that clinical reasoning can be taught, but it requires a structured and systematic approach that incorporates consistent feedback and self-reflection. Schuwirth (2002) argued that teaching clinical reasoning in the classroom alone is not sufficient, as it must be learned in the "context to which it is applied" (p. 696). This view is in alignment with the theory of situated cognition, which argues that "the activity in which knowledge is developed and deployed…is not separable from or ancillary to learning and cognition" (Brown et al., 1989). Brown, Collins, and Duguid (1989) argued that cognitive apprenticeship, which supports learning in a context-specific domain, allows experts or teachers to "promote learning, first by making explicit their tacit knowledge or by modeling strategies for students in authentic activity" (p. 39). Since clinical reasoning is a series of complex, tacit thought processes, cognitive apprenticeship could be a promising method to allow more experienced physicians to explicitly share with learners how they reason through clinical problems, particularly for the analytic, system 2 processes, hypothetico-deductive and scheme-inductive reasoning.

The review so far is intended to provide an overview of the clinical reasoning process, or processes, and their relation to reasoning and learning concepts from the education and cognitive science literature. While two suggestions for teaching clinical reasoning, cognitive apprenticeship and deliberate practice, are briefly shared above, they by no means represent an exhaustive list of pedagogical approaches to clinical reasoning. I will not dive deeper into exploring methods for teaching clinical reasoning in this review, but offer the ones I have to reinforce that clinical reasoning is a skill that can be developed (Higgs et al., 2018). Having this understanding of what clinical reasoning is and that it can be improved upon with training will help provide context for the forthcoming discussion on assessment approaches of clinical reasoning.

Assessment of Clinical Reasoning

While clinical reasoning has been studied for many years, educators still debate over the best approach to assess clinical reasoning (Durning et al., 2013). Both general education and medical education research offers valuable guidance for educators and practitioners to follow in the creation and use of assessments. In 2018, leaders in medical education assessment endorsed that a "good" assessment should have 7 characteristics: validity or coherence; reproducibility, reliability, or consistency; equivalence; feasibility; acceptability; educational effect; and catalytic effect (Norcini et al., 2018). The final two components relate to a primary stakeholder involved with assessment: the learner. If an assessment has educational effect it "motivates those who take it to prepare" (Norcini et al., 2011). Conversely, assessments with a catalytic effect enhance future learning. Workplace-based assessments are one type of assessment that can have strong educational and catalytic effects because they are centered around real-life practice (Norcini & Burch, 2007). Assessing students in the environments in which they will be expected to practice once they become physicians provides opportunities for meaningful feedback that students can use as they strive to become experts.

Any assessment can be classified as either formative or summative. Summative assessments are completed usually at the end of a course or experience to summarize student performance, often with the intention of assigning a grade or decision for promotion (Pangaro, 2012). Most summative assessments are high-stakes exams, exams whose outcomes have potentially serious consequences like grade promotion, certification, or licensure (Ryan, 2002). For this reason, any high-stakes assessment should have an internal consistency coefficient, Cohen's alpha, of 0.90 (Bandalos, 2018) to ensure it is accurately assessing the learner. In UME, the United States Medical Licensing Exam (USMLE) Step 1 and Step 2 exams, are the most prominent high-stakes exams as they determine, in many schools, whether a student can progress through the curriculum or graduate. Formative assessments are intended to help the learner identify areas of weakness and to have a catalytic effect, or stimulate more learning (Norcini et al., 2018). These assessments are used in low-stakes settings, where the outcome of the evaluation does not have significant consequences on the progression of the learner. Examples of a low-stakes evaluation in UME include quizzes during the didactic curriculum and mid-point evaluations on clerkships (clinical learning experiences).

Having a balance of formative assessments (assessments *for* learning) and summative assessments (assessments *of* learning) is essential to creating an effective program of assessment in any school or curriculum A program of assessment allows for "the whole picture of a student's competence to be obtained by a careful selection of assessment methods" (Schuwirth & Van der Vleuten, 2011). A mixture of frequent formative assessments and summative assessments can help ensure a learner is on track and the curriculum is effectively meeting the learners' individual needs (Pangaro & ten Cate, 2013).

In 2019, Daniel et al. published a systematic review of clinical reasoning assessment methods. Their sweeping review ultimately included nearly 400 articles specifically focused on clinical reasoning in medical learners/professionals (i.e. medical students, residents, and physicians). In the review, the authors offered a framework for identifying the multiple steps in clinical reasoning and how to assess each phase of the process: information gathering; hypothesis generation; problem representation; differential diagnosis; leading or working diagnosis; diagnostic justification; and management and treatment (Daniel et al., 2019) and classified existing assessment methods into three categories: non-workplace-based assessments, simulation-based assessments, and workplace-based assessments. I will explore each of these categories below.

Non-Workplace-Based Assessments (NWBAs)

The majority of methods identified by Daniel et al. (2019) were non-workplace based assessments (NWBA). These assessments offer a high degree of reliability and standardization, making them appropriate for high-stakes assessments, however, they do not take into account all the phases of the clinical reasoning process (Daniel et al., 2019). Not being representative of the entire reasoning process that occurs in practice, one can make limited assumptions about learner progression using NWBAs when used in isolation.

Most NWBAs are written exams administered either electronically or by paper. Multiple choice question (MCQ) exams were the most common NWBA reviewed (Daniel et al., 2019). These exams often require students to make a diagnosis based on a clinical vignette or scenario. Some exams take this one step further by presenting the students with additional information and asking them to select a new diagnosis (or their original) based on the new information (Kelly et al., 2012). Other written exams are essay format or free-form asking students to outline their thought processes using words or mental maps (Daniel et al., 2019). Others have explored using verbal exams to employ think-aloud (TA) protocols to assess how students solve reasoning problems as they are doing so (Surry et al., 2017). In research, TA protocols have proven useful in helping understand problem-solving skills without altering the examinees thought processes (K. A. Ericsson & Simon, 1998). Interestingly, when the results of the TA studies were compared to a written, MCQ exam that tested similar scenarios, the results were aligned (Surry et al., 2017). Daniel et al. (2019) argue that TAs, when used for assessment, should be used primarily in low-stakes assessment settings due to the subjectivity of the assessor and significant amount or resources it takes to train, administer, and score the exam.

During development, non-MCQ exam NWBA formats are often compared with existing, high-stakes validated MCQ exams that are used nationally. Although MCQ exams are the most likely method to achieve high reliability (Daniel et al., 2019), I argue that these exams often rely mostly on a student's ability to recall biomedical knowledge, ignoring other important facets of the clinical reasoning process, such as problem representation and diagnostic justification. So, while some NWBA formats (including MCQs) have the potential to provoke thought responses similar to those that physicians use in clinical encounters (Surry et al., 2017), they cannot be used to assess clinical reasoning alone. This conclusion is in alignment with the traditional education and expertise literature that argued that MCQ exams are not comprehensive enough to capture the entire clinical reasoning process (Ennis, 1993).

Simulation-based Assessments (SBAs)

In response to the lack of comprehensive assessment provided by MCQ exams and other NWBAs, objective structured clinical exams (OSCEs), and later simulation activities, were developed to help mirror scenarios that would occur in the clinical setting. These two forms of assessment were identified by Daniel et al. (2019) in their systematic review as simulation-based assessments (SBAs).

Unlike in an MCQ exam were students are given the information needed to form a diagnosis, OSCEs require students to collect the information from an actor, or standardized patient, who has been provided a script. The encounter is carried out like a traditional clinic office visit. Students are then scored using a checklist on six domains, including information gathering and diagnosis. The OSCE checklists are commonly completed by the standardized patients, as this requires less time from expensive, busy faculty. However, some OSCEs use faculty raters, who watch the encounter via video from a separate room. These raters will complete the checklists themselves, and may also provide additional comments or global ratings of the student's performance.

One of the first studies using OSCEs was a large-scale performance-based examination using standardized patients in a simulated clinical environment (Vu et al., 1992). Similar to the NWBAs described above, Vu et al. (1992) also used a nationally administered MCQ exam to assess the validity of an end-of-clerkship OSCE and a generalizability study to assess the exams reliability across five years of student performance data. The correlation between the performance on the MCQ and the overall score on the exam ranged from 0.20 to 0.65 across the five years, showing a weak to moderate correlation. The authors argued that while not ideal for high-stakes decision, since the exam was still in early stages of development it was acceptable for the low-stakes setting in which it was applied (Vu et al., 1992).

SBAs are an attractive method of assessment because of their ability to mirror real-life practice while providing a standardized assessment environment that can be calibrated to the level of the learner. For this reason, they have become a popular high-stakes exam format (Norcini et al., 2018) and represent the "best-case" scenario for learner performance (Daniel et al., 2019). This is because they rule out contextual factors that can alter student performance, such as patient-specific factors and the clinical setting (Durning et al., 2013). However, SBAs incur a high-cost to the administrating institution due to the need to hire actors to serve as standardized patients or to purchase and maintain the technology required for simulation activities. Despite these limitations, these exams are used widely in UME programs (Daniel et al., 2019).

Workplace-based Assessments (WBAs)

The final category, and arguably the most intuitive, identified by Daniel et al. (2019) is workplace-based assessments (WBAs). The advantage of WBAs is that they assess students in the environments in which they will be expected to practice in the future. The most common WBAs include direct observation, global assessments, and written notes (Daniel et al., 2019), which I will explore in further detail below.

To aid with director observation, medical educators have relied on and developed broader-level frameworks when assessing learners in the clinical setting. These frameworks include: Bloom's taxonomy (remembering, understanding, applying, etc.); Miller's triangle (knows, knows how, shows how, does); RIME (reporter, interpreter, manager, educator);

entrustment scales, and others (Pangaro & ten Cate, 2013). Often these frameworks are translated into scales or checklists that allow evaluating faculty to turn their observations of students into systematic interpretations of the students' ability levels (Norcini & Burch, 2007). These interpretations are submitted via electronic or paper forms that can be translated into global assessments, which give students a score or grade that can be used to assess readiness for progression (Daniel et al., 2019). These tools work well for the information gathering, differential and leading diagnoses, and management and treatment phases of clinical reasoning, but are not always granular enough to assess a student's thought processes (i.e. hypothesis generation and diagnostic justification) (Daniel et al., 2019; Norcini & Burch, 2007). Furthermore, depending on the tool, global assessments may be able to assess overall competence, but fail to identify deficiencies in specific milestones, which are "observable tasks that combine knowledge, skill and attitudes" (Pangaro, 2012, p. 9). For example, an individual can be identified as a "reporter" in the RIME framework if they can gather a sufficient history and share that information with others, but the way that they collected their information may have been unorganized, which could affect their ability to collect the correct information in the future when confronted with more complex cases.

Examining patient notes written by a student can offer a window into that student's thought process as written notes contain many of the clinical reasoning components identified above. Medical students are expected to write notes on many of the patients they see during their clinical clerkships. Depending on the setting (inpatient vs. outpatient) and medical specialty (e.g. Internal Medicine, Pediatrics, etc.), the note template will vary, but common components require students to document the information they have gathered from both the interview and physical exam, articulate a diagnosis, and justify why that diagnosis has been determined given the information they collected. More advanced learners must then also provide a recommendation for a management and treatment plan (Daniel et al., 2019). Specific note templates can also be useful in providing structure for assessors to see a student's thought process, such as those that specifically ask students to rank diagnosis and treatment plans based on probability.

Workplace-based assessments are attractive assessment methods because they allow educators to identify how a student is performing in the real-life clinical setting, where patients do not adjust their illnesses and complexity to the level of the learner. The same reason they are attractive, however, makes them less attractive for high-stakes exams since the conditions cannot be controlled and there is no way to ensure difficulty is standard across examinees. Norcini, Blank, Duffy, and Fortna (2003) and Ansari, Ali, and Donnon (2013) found that a minimum of 12 observations are required to achieve appropriate reliability measurements for advancement decisions (as cited in Daniel et al., 2019). Lastly, since work-place based assessments are highly subjective, there is also a high potential for bias among raters. Training for assessors (which can be any clinical faculty member within an institution) is necessary to ensure faculty are aware of any implicit biases they may have towards learners and to ensure calibration of global assessment scores (Daniel et al., 2019; Norcini & Burch, 2007).

To make an assessment decision with the data gathered during these commonly used WBA methods, a faculty member must interpret the data and make a judgement about a student's performance. Although comprehensive and robust, Daniel et al. (2019) did not address a critical component in the assessment of clinical reasoning: the clinical reasoning steps identified are not visible to assessors in practice, nor how they manifest in medical students or novice learners. I argue that without this information medical educators lack clear criteria to make assessment decisions. This is further complicated by the fact that most faculty are far removed from employing the same clinical reasoning processes that they are assessing, since medical students use a slower, analytical approach to clinical reasoning than expert physicians. Without a clear framework on how to make assessment decisions, I argue that faculty rely on their own intuition rather than robust criteria. This affects the quality of feedback that a student receives and can result in faculty making unreliable advancement decisions.

How Faculty Assess Learners

Work to understanding how medical faculty complete assessments is not new in medical education, but limited work has been done in the medical school space, nor have I found work that specifically relates to the assessment of clinical reasoning. In effort to understand how faculty make assessments of residents via direct observation, Kogan et al. (2011) conducted qualitative interviews with faculty from a number of east coast medical schools after the watched a number of pre-recorded and live encounters of residents with standardized patients. Participants were asked to complete a common rubric used with in graduate medical education, the mini-CEX, for each encounter and then asked to complete a 15-minute semi-structured interview. Using grounded theory, Kogan et al. (2011) found a number of factors that influence participants assessment decision-making, most notably that participants commonly make inferences or assumptions about the learners and rely on three types of frames of reference: self, others, and standards of care.

In 2014, a group of international medical education researchers published a paper on common themes and implications of assessor cognition based on findings in their own research and practice (Gingerich et al., 2014). The group noted several key perspectives: first, that the assessor is trainable, and therefore their behavior is potentially modifiable through training or faculty development; second, that the assessor is fallible and not immune to error or limits on their cognition; and lastly, that the assessor is “meaningfully idiosyncratic” and that their individual differences will inevitably emerge in their assessment (Gingerich, et al., 2014). These findings have interesting implications specifically on the assessment of clinical reasoning. Because clinical reasoning is a cognitive process, faculty may be more likely to make inferences, rather than basing their assessments on observable behaviors or statements. Faculty members are also not machines and can be prone to cognitive bias and errors like any other individual. Young et al. (2014) found “[Cognitive load theory] has particular relevance to medical education because the tasks are complex and may impose a cognitive load that surpasses the [working memory] capacity of the learner” (p. 371). In assessment, cognitive bias may be further amplified if the assessor is performing this complex task of clinical reasoning on a patient while simultaneously assessing the student. Gingerich et al. (2014) also noted, “rather than ‘objectively’ recalling what they have just observed, people may unconsciously ‘fill in the blanks’ based on what their stereotypical beliefs suggest” (p. 1059). This may also mean that if asked to provide assessment far after they actually observe the student, their assessments could be swayed.

The final step to understanding how faculty assess learners is understanding how they share their findings with students through verbal or written feedback. A commonly accepted best practice in the provision of feedback to medical students and residents is that it should occur first face-to-face before being submitted in writing. Unfortunately, in practice, this does not always occur and students are only left with a written narrative on how they performed. One study examining written feedback provided to residents by attendings rated the overall quality to be low to moderate (Jackson et al., 2015). Other studies have also shown narrative feedback to be influenced by gender and racial/ethnic bias (Rojek et al., 2019). Ramani and Krackov (2012) published an article providing twelve tips for giving effective feedback, however, the focus was on delivering face-to-face as opposed to narrative feedback. If we look outside the medical education literature, traditional education research suggests that feedback should address three components: where is the learner going, how are they doing, and where are they going next (Hattie & Timperley, 2007). This can be done by addressing the quality of the learner's work, the process the learner used to create the work, or personal characteristics that may be inhibiting their progress (Hattie & Timperley, 2007). Although discussed by many medical schools, very little is available on what constitutes good *written* feedback in the medical education research.

In this dissertation, I build upon the work of Kogan et al. (2011) and Gingerich et al. (2014) to try to understand how faculty members make clinical reasoning assessment decisions in the moment, as opposed to at the end of an encounter. I believe this is an important distinction because it will hopefully allow us to understand the tacit, cognitive processes faculty members use when making assessment decisions, as opposed to relying on their own recall of what they were thinking, which we know can be influenced and altered the further they get away from the initial thought (K. A. Ericsson & Simon, 1998). In order to truly understand how faculty make these decisions, I examined how faculty make assessment decisions by analyzing the thoughts they have while presented with an assessment opportunity. The method I used to do this is through the use of think-aloud interviews.

Think-Aloud Interviews

Think-aloud interviews (TAs) have been widely used in the cognitive science and medical education fields to allow researchers to understand how individuals approach problem-solving tasks. The methodology of TAs, and the systematic approach to their analysis, protocol analysis, were honed by Ericsson & Simon (1993) and have proven useful in helping understand problem-solving skills without altering the examinees thought processes (Ericsson & Simon, 1998). In health professions education research, TAs have primarily been used to understand how medical students and expert clinicians approach clinical decision-making (Funkesson et al., 2007; Lundgrén-Laine & Salanterä, 2010). More specifically in medical education, researchers have used TAs to understand the different ways in which novice and expert medical practitioners make diagnoses (Coderre et al., 2003) and have paired them with concept mapping to identify if students are using inductive or deductive approaches to reasoning (Pottier et al., 2010). Medical educators also used TAs to understand how students approach answering MCQ exams that are based on medical knowledge and reasoning ability (Daniel et al., 2019; Surry et al., 2017). All of these studies employed concurrent TAs, in which participants verbalize their thinking *while* completing the assigned task, rather than retrospective TAs, when participants reflect on their thinking *after* a task is completed (Fonteyn et al., 1993). This distinction is important because the further removed a participant is from the activity of interest, the more likely it is for the data to be influenced by time to reflect and interpret how they thought about something rather than the thoughts themselves (K. A. Ericsson & Simon, 1998).

Using TAs to assess a student's thinking is not new to medical education, however, to my knowledge TAs have not yet been used upon faculty to understand how they are assess medical students. In other higher education settings, researchers have conducted TAs to identify how professors grade university students (Boyd et al., 2009). In 2019, a study using narrative comments from an OSCE in a pharmacy school were given to faculty members who were asked to do a TA as they interpreted the comments (Wilby et al., 2019). In this work, the authors found that assessors arrived at similar conclusions (e.g. pass vs fail), but how they came to their conclusion varied. In addition, only the narrative comments about students' communication abilities from each station of the OSCE were presented to the TA participants, instead of capturing TAs from faculty observing students directly.

Using TAs for Faculty Reasoning

Think-aloud interviews offer the researcher rich qualitative data; data that allows the researcher to understand how things work or why an individual made the decisions that he or she made. Ericsson and Simon (1993) provide clear guidelines on how TAs should be analyzed using protocol analysis, a tightly structured method for analyzing TA data once the data has been transcribed. However, protocol analysis is also dependent on using a deductive analytical approach based on already existing theory. Since there is currently no existing theories on how faculty make assessment decisions about clinical reasoning, following a strict protocol analysis is not appropriate. Gu (2000) argued that when studying a new phenomena, which requires an inductive approach, using TA data, a grounded theory approach may be the only option (Gu, 2000).

Grounded theory differs from other qualitative methods in that its aim is to build a theory where one currently does not exist (Merriam & Tisdell, 2015). As mentioned above, because no theory currently exists for faculty reasoning, this approach seems appropriate. However, since my attempt is to merely understand how faculty make assessment decisions, and not to build a theory or framework around faculty reasoning entirely, a pure grounded theory approach is arguably not appropriate. Thematic analysis offers a more flexible approach for inductive qualitative work (Braun & Clarke, 2006). Similar to grounded theory, thematic analysis employs the use of codes to develop broader themes about the data, without necessitating those themes to be used to create a theory. Using a thematic analysis of the think-aloud interviews collected from faculty as they make assessment decisions allowed me to capture themes across faculty, which helped me better understand how faculty make assessment decisions and create a model of the faculty reasoning process.

III. Methods

Participants

While there is no clear consensus on how large of a sample is necessary to conduct a successful TA study, both Ericsson and Simon (1993) and Leighton (2017) highlight the importance of selecting study participants that reflect characteristics and expertise of the larger population. For this reason, I aimed to recruit 15 faculty members who vary in the following demographics: gender, age, years in practice, years actively engaged in medical education, and clinical specialty. Any faculty member who holds an M.D. or D.O. degree and is practicing clinical medicine was eligible to participate.

In late-October 2020, a handful of University of California, Davis School of Medicine faculty members were e-mailed and asked to participate in a study about the assessment of clinical reasoning in medical students. These faculty members were selected based on their experience in medical education and my experience working with them on education-related projects in the past. Within the email (see Appendix A), the faculty were also encouraged to pass along the email to any of their interested colleagues. Three additional follow-up emails were sent to educators in the Surgery, Ob/Gyn, and Emergency Medicine department in hopes of soliciting broader representation from those specialties.

Fifteen faculty educators responded back indicating interest and willingness to participate. Two fellows also reached out and asked if they were eligible, but were declined due to the study being focused on the assessment decisions of *faculty*. Of the 15 faculty members who indicated interest, all but one completed the study (due to lack of response when trying to schedule). No incentives were offered for the participants' participation. A summary of the participant characteristics can be found in Table 1.

Table 1

Participant characteristics

	n	%
Gender		
Male	6	43%
Female	8	57%
Specialty		
Emergency Medicine	2	14%
Family Medicine	1	7%
Internal Medicine	7	50%
Ob/Gyn	1	7%
Pediatrics	1	7%
Psychiatry	1	7%
Other	1	7%
Years as Faculty		
5 years or less	6	42%
Less than 3 years	3	21%
3 – 5 years	3	21%
More than 5 years	8	58%
6 – 10 years	1	7%
More than 10 years	7	50%

Materials

The stimulus for the TA was a recorded encounter from the summative Clinical Performance Examination (CPX), a required, eight-station OSCE. Each station, or case, is no more than 20 minutes and is scored using pre-determined checklists which are completed by the SP immediately after the encounter. In consultation with the UC Davis CPX director to find a case that was constructed in a way to be able to assess clinical reasoning, an abdominal pain case was selected and an additional interstation was added. This interstation asked the student to leave the room briefly to collect their thoughts and return to provide an oral presentation for a new actor who played the role of the student's attending.

Once the exam was completed and scored, I identified which video to use by narrowing down the pool to students who a) consented for their videos to be used for education or research, and b) scored within 0.5 SD above or below the mean on the History Taking and Physical Exam checklists and Overall score for both the exam and the individual case. Once I identified an individual student I requested the CPX director to verify my selection as an appropriate choice for asking faculty to assess clinical reasoning, which she did. The video recording of the student's encounter with the standardized patient and oral presentation became the stimulus for the TA.

In the past, this exam occurs in the month following the completion of their clerkship year (May) and both the student and standardized patient (SP) are in the room. However, due to COVID-19 related restrictions placed on in-person educational activities beginning in March 2020, the exam was postponed to August 2020. In addition, the exam was reconfigured to be a virtual encounter, that is the student was engaging with the SP via Zoom. This means that while the student was recorded for the encounter, the recording did not include clear video of the SP, just clear audio. This also had significant implications for the physical exam portion of the exam. Unable to conduct a true physical exam on the patient, students were instructed to describe what maneuvers they would do and were provided with any sensitive exam findings in the Zoom "Chat" function. Because this altered the way in which a student could also collect physical examination information, participants were alerted that they would receive the relevant physical exam findings via the "Chat" function in Zoom (Appendix B).

Study design and data collection

Think-aloud interviews (TAs) took place between November 2020 – January 2021. Each interview lasted for one hour and was scheduled according to the participant's availability. Due to the surge of COVID-19 cases in the Greater Sacramento area that occurred during this time frame, participants completed TAs via Zoom, which also allowed for recording of the interview and the automatic creation of an interview transcript.

At the start of the interview, participants heard the following script to orient themselves to the interview process:

> Thank you for agreeing to participate in this study on clinical reasoning assessment. For the purposes of this study, clinical reasoning is defined as "a skill…wherein clinicians observe, collect, and interpret data to diagnose and treat patients." I'm interested in understanding how faculty make assessment decisions about medical student's clinical reasoning abilities. To do so, I'm going to employ a think-aloud interview process with a video-recorded standardized patient clinical encounter. This process asks you, the participant, to vocalize any thoughts you have as you review the video. When I say any, I mean ANY, with a specific focus being on assessing the student's clinical reasoning ability. Because you will be watching a video as you do this, you will have total control to pause the video using the spacebar at any point to say whatever you are thinking. I understand that this may feel awkward, but try to pretend I'm not here. The only time I may interject is if you become silent for too long. In which case, I will gently prompt you by saying, "Keep talking" or "Any thoughts?" At the end of the video, I will also ask you to provide brief summative feedback, throughout which I would also like you to continue to think-aloud. Please keep in mind that there is also no right or wrong answers throughout this entire process and you do not need to justify why you are thinking whatever you are thinking. My only goal is to learn about how you assess medical students. [We will also have the opportunity to practice the think-aloud interview process]*, but before we get there, do you have any questions?
>
> We will now begin watching the video of the encounter. Due to COVID, this is a virtual encounter with a standardized patient, Hannah Wesley. I've provided the Door Notes visible to the student before the enter the room in the chat.
>
> * For the first participant I attempted to show a video from YouTube to orient the faculty member to the think-aloud process. Doing the interview via Zoom and not having direct control over the video made this challenging and ineffective, so the practice-run was removed for future participants

Because the TAs were conducted via Zoom, I had to experiment with different options for allowing the participants to pause the recording to express their in-the-moment thoughts. Prior to the first interview, I practiced doing this two ways with a colleague. The first way was to give the participant remote control of my screen so they could pause the video themselves and vocalize their thoughts. This method seemed to work well when internet connectivity was not a concern and bandwidth was high. However, once I started recording, it was much more difficult for individuals on the other side of the meeting to do this. The other option was for me to pause the video based on cues from the participant (i.e. hand raise, start talking). This worked more consistently with seemingly minimal disruption to the participants ability to convey their thinking.

At the end of the encounter, participants were asked to provide summative comments on the students clinical reasoning ability in an open text box, similar to how they would in MedHub, the UC Davis School of Medicine's online assessment system that faculty regularly use to assess medical students. The prompt for the summative comments was:

> "**Summative Comments**: Please provide feedback to the student on their clinical reasoning ability.
> **Reminder:** Clinical reasoning is the ability "to observe, collect, and interpret data to diagnose and treat patients."

The comments were collected via a Qualtrics survey that only I had access to. Participants were instructed to continue to think-aloud as best they could while providing their summative comments and were recorded throughout to ensure the capture of any additional thoughts. This was all considered part of the think-aloud transcript. The summative comments were extracted from Qualtrics and analyzed separately from the think-aloud transcripts.

Once their comments were completed, within the same Qualtrics survey faculty members were also asked "Please indicate approximately how many years you've been involved in teaching, assessing, or working with medical students as a faculty member." The options included: Less than 3 years, 3-5 years, 6-10 years, More than 10 years. This information was added to the participant characteristics I already had from past interactions with the participants (including gender and specialty).

During each TA I also took hand-written notes of things that popped out to me as being specifically related to clinical reasoning or interesting. These notes were transcribed and uploaded to the qualitative analysis software as analytic memos linked to each participants TA.

All video, audio recordings, and transcript outputs were uploaded to a secure OneDrive folder. Data from the Qualtrics survey containing the summative comments and "years as faculty" question were exported and also saved in the secure OneDrive folder. While the transcripts were mostly complete using the Zoom auto-transcription feature, a significant amount of cleaning needed to be done to separate the student and SP dialog from the participants thoughts and correct imperfect transcriptions. All cleaning of the transcriptions was done in Microsoft Word. Once cleaned, the transcriptions and summative comments were imported into Dedoose (version 8.3.45).

Analysis

The transcripts and summative comments were analyzed using thematic analysis to identify themes within the data. Braun and Clarke (2006) define thematic analysis as “a method for identifying, [analyzing] and reporting patterns (themes) within data.” More flexible than traditional qualitative methods, thematic analysis can be used inductively or deductively on any level of depth of analysis. Grounded theory is similar to thematic analysis in that both identify codes and use those codes to build larger themes, however grounded theory uses a constructivist framework to build a theory where one currently does not exist (Merriam & Tisdell, 2015). Thematic analysis allows for a similar methodology without requiring a theory to be created (Braun & Clarke, 2006). As no theory exists on how faculty assess clinical reasoning specifically, thematic analysis allowed for a more exploratory, inductive approach. The analysis process followed the five steps outlined by Braun and Clarke (2006): familiarize yourself with the data; generate initial codes; search for themes; review themes; and define and name themes.

Familiarize yourself with the data & Generate initial codes

Because I conducted all of the interviews, took notes on all of the interviews, and cleaned up all the transcriptions, I felt very familiar with the data prior to beginning my analysis. Therefore, I merged steps 1 and 2 and started generating initial codes once the data was imported into Dedoose. A code is "a word or short phrase that symbolically assigns a summative, salient, essence-capturing, and/or evocative attribute for a portion of language-based or visual data" (Saldaña, 2016, p. 4). In my coding, I used primarily initial codes and concept codes (Saldaña, 2016). After my first pass of coding each transcript and summative feedback I drafted an analytical memo of what I noticed, codes that I used or created, questions that were arising, and any initial patterns that were emerging (Deterding & Waters, 2018). Additional codes were created based on seven steps of clinical reasoning published by Daniel et al. (2019): information gathering; hypothesis generation; problem representation; differential diagnosis; leading or working diagnosis; diagnostic justification; and management and treatment.

Search for themes & Review themes

After coding each transcript and summative comment and writing each analytic memo I reviewed each memo and drafted a list of questions that emerged from my first pass. While my initial research questions were fairly broad, after doing this first pass of open coding, I was able to create more discrete research questions that helped refine my codes and helped to lump codes into categories (Saldaña, 2016). The list of questions and explanations can be found in Appendix C. These questions also prompted me to go back to the literature, specifically to review the work done by Kogan et. al. (2011) on how faculty assess resident workplace-based observations and identify past research around cognitive load for assessors in medical education. This review provided useful language to clarify some of the earlier codes I had created and start noticing similar themes in my own work.

After reviewing the literature and reflecting on the questions, I reviewed my codes and tried to identify themes by collating the codes under "parent codes", which I used to organize my codes in Dedoose into categories (Merriam & Tisdell, 2015). Once grouped into categories, I took all the categories and codes and mapped them out using Post-it notes and easel paper (see Figure 1).

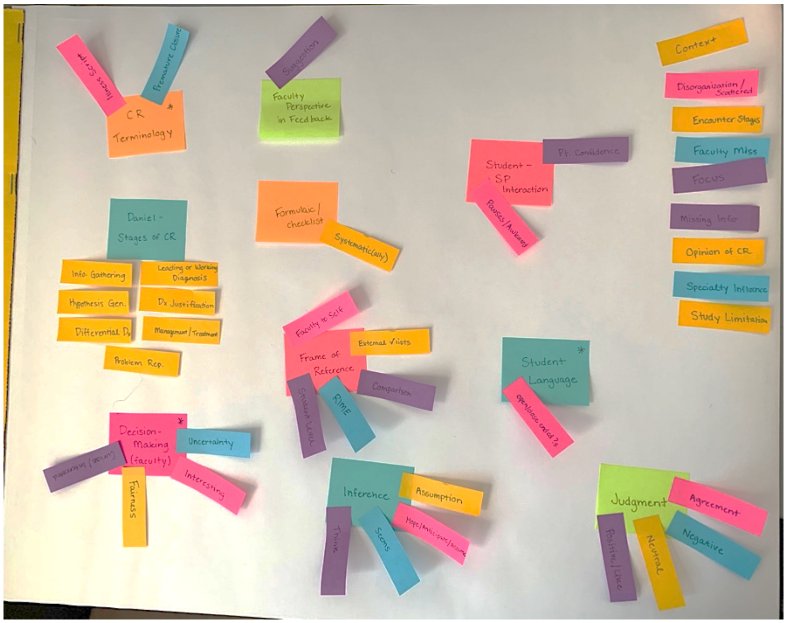


Figure 1. Codes in categories, with some uncategorized codes (version 1)

I then took all the uncategorized codes and put them on a separate easel sheet to try and see if there were any connections between them or if they stood alone. I realized I was able to create a new category for some and assign the remaining codes to already existing categories. Once all the codes were in a category, I looked to see if any of the codes within each category were repetitive. This brought me back to the data to reassess how the codes were used, provide more specificity, and clean up the codes. This resulted in the creation of new codes or the merging of others.

Define and name themes

As the categories became clearer, I started lumping together categories, which ultimately resulted in the identification of the following themes: assessor factors, student factors, situational factors, integration, and judgment. In addition to these five themes, the codes collected in the analysis of the summative comment data were lumped under a "feedback" category. A theme is "an extended phase or sentence that identifies what a unit of data is about and/or what it means" (Saldaña, 2016, p. 199). Identifying themes and seeing the ways they interact, led to further refinement of my codes and categories. This iterative process repeated itself several times until all categories had no excerpts attached to them and each category and code was unique and mutually exclusive (Merriam & Tisdell, 2015). The final grouping of categories and themes can be seen below in Figure 2. A final list of codes with definitions can be found in Appendix D.

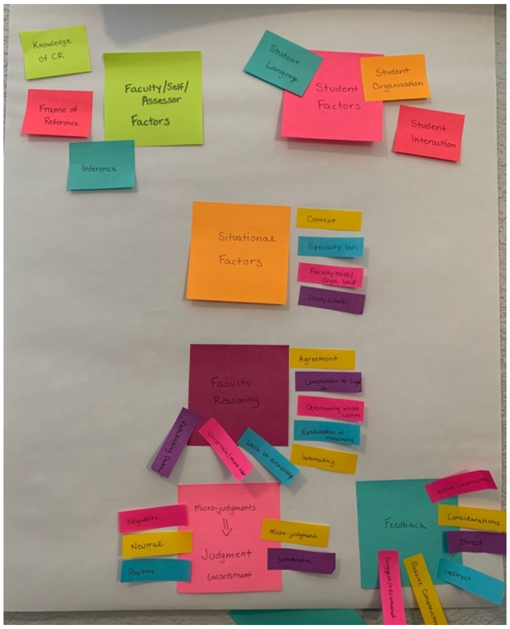


Figure 2. Final representation of categories and themes.

IV. Results and Discussion

After coding the transcripts and organizing the codes into categories, I identified six themes in the data: *assessor factors, student factors, situational factors, integration, judgment, and feedback*. Once the themes were identified, I began to see similarities between the themes and how they are interconnected with one another. For example, *student factors* and *situational factors* describe different components of the encounter, either those related to the student or the specific case, respectively. Both of these factors are observed and interpreted by the faculty, each of whom uniquely come into the assessment scenario with their own knowledge, opinions, and interpretations. These are the *assessor factors*. How the faculty member integrates these different factors together and processes them is captured in the *integration* theme. The decision-making process (*integration*), ultimately leads to *judgments* of clinical reasoning ability. These *judgments* are then expressed to the student via some sort of feedback mechanism, which for this study was narrative comments provided at the end. I have called this entire process the faculty reasoning process. The interactivity of the themes (which will be noted in *italics* for the remainder of the chapter) within this process is described in more detail in the following section. The remainder of this section focuses on the results from the coding and theme identification process.

General Findings

The think-aloud interviews (TA) transcripts varied in length from approximately 21,000 characters to approximately 60,000 characters. Not every code was applied to every transcript, however once the codes were categorized, every category and theme was within each transcript, with the exception of *situational factors*. When appropriate, codes were applied multiple times to the same transcript in different segments. The most commonly applied codes were "positive" and "negative," which were in reference to faculty statements about something they liked or did not like about what the student did in the encounter. The next most commonly applied codes were "differential diagnosis," when a faculty member referred to the student's differential, "the process of differentiating between two or more conditions which share similar signs or symptoms" (Oxford English Dictionary, 2020), and "faculty to self," which I defined as when a participant expresses what they would do in the scenario or would have liked to have seen the student do - comparing the student to themselves. A full list of codes, their frequencies, and their definitions can be found in Appendix D.

The data for the summative comments is much smaller (each ranging between 492 and 1328 characters) than the transcript data from the think-aloud interviews. The summative comments data is much more relevant to the *feedback* theme that will be discussed in the last section of this chapter. For this reason, all future tables up to that point will share only data from the think-aloud transcripts.

The text above provides an overview of the categories and themes and how they were noted within the TA transcripts. In the following sections I dive more deeply into each of the themes and categories and include individual level codes within the tables. I chose to include this level of detail to clarify how I analyzed the transcripts (what codes I used) and how those codes relate to one another.

Assessor Factors and Integration

Assessor Factors

The first theme, *assessor factors*, captures the different internal contributions the faculty is making from their own experience or understanding to the assessment of students. This theme contained the following three categories: knowledge of clinical reasoning, frame of reference, and inference. The codes and categories used to identify this theme were well-represented across all the participant transcripts.

Frame of Reference

I borrowed the term "frame of reference" from the work of (Kogan et al., 2011) in identifying how faculty assessors assess residents. Kogan et al. (2011) identified three frames of reference: performance by oneself, the performance of other doctors, and a standard of performance considered to be necessary for patient care. Similarly, "frame of reference" in this study refers to the measures or frameworks participants used to compare the student's performance against some external metric, however, with slightly different reference points. The differing frames of reference identified include faculty to self, student to others, and use of external frameworks.

The first reference point closely aligns with past research in that participants often compared what the student was doing to how they would have done it in the encounter. When describing the student's approach to data gathering, one participant commented:

> What I like to do is kind of a more focused review systems on kind of like the abdominal and GU region first. So I would have carried along with the thoughts [they] had about like nausea, vomiting, diarrhea, constipation. I might have like pulled it further us like any darker or tarry stools any you know bright red blood,

> you know acid reflux, things of that nature… Because it like, although [they] did preface like these are tangential questions, I think there's still a way of kind of like organizing it and focusing it too. (Participant 10)

The quote above is referencing how the participant would have organized the interview differently than the student, which is very relevant to the assigned task of assessing clinical reasoning. In other excerpts, participants expressed their own preferences for much smaller differences (i.e., referring to the student as “woman” instead of “girl”). These smaller details were often a reflection of the participant’s preference, as opposed to commenting on something the student did objectively wrong, and were not always related explicitly to the student’s clinical reasoning. I find this interesting to consider because while seemingly small, concordance, or lack thereof, with how the participant would have conducted the interview could color their assessment of the student’s performance. If each faculty member comes into an assessment using themselves as the benchmark for an assessment, the criteria for performing well cannot be clearly explained to the student and it could constantly change with each new assessor if their preferences differed from one another. This raises questions about validity and fairness.

The second frame of reference used by participants was attempting to place the student at a certain level of training by comparing them to other learners they worked with in the past in order to help them calibrate their assessment. The estimates shared by participants ranged from first-year medical student to resident and often changed as the encounter progressed. One participant commented, “[The student] did an adequate assessment and an adequate history with the patient, including the pelvic exam in a way that I would expect like a family medicine intern to do or an internal medicine intern to do” (Participant 4). Another commented, “You know it's pretty typical for third year students, if that's what [the student] is” (Participant 6).

Since participants were not given any sort of rubric or scale to use when assessing the student’s clinical reasoning ability and were not told the student’s level of training, I am not surprised that faculty sought out other reference points to aid them in their assessment. Several participants asked mid-encounter if they could be told what year the student was in medical school, but I did not tell them until the end of the interview. Faculty in the School of Medicine are used to working with residents and medical students of varying levels in the clinical setting, and although are provided with behaviorally-anchored rubrics to assess students, it is generally believed that most rely on comparison of students to others in determining how to grade students for coursework. Some faculty were able to adjust their classification of the student continuously throughout the encounter as they received new information of the student’s performance. For example, Participant 14 noted the student could be at a resident level early on in the encounter, but shifted their assessment downwards to being an advanced medical student by the end. This shifting and changing of assigning student level will be discussed further in the *judgment* section.

The last frame of reference, the use of external frameworks or assessment systems, was used less commonly by all the participants. The most common external framework noted was the RIME (Reporter-Interpreter-Manager-Educator) framework (Pangaro, 1999). One example of a participant using the RIME framework to assess the student is when Participant 11 stated:

> I think you've probably heard the framework of the RIME. The reporter, the interpreter…So I think he's like working towards being a reasonable reporter and using a template that seems reasonably complete, but it doesn't seem like there's any reasoning being applied to the way he is using that template. (Participant 11)

One could argue that using an external rubric or scale (such as RIME) demonstrates a more advanced approach to assessment, as it suggests the use of criterion-referenced as opposed to normative-referenced (comparing the student to other students) grading. The higher frequency with which some participants referred to these frameworks, also could imply that they constantly had them in the back of their mind and revisited them more frequently when assessing the student's performance throughout the encounter.

Frames of Reference Summary

The identification of these different frames of reference (self, others, and external frameworks) demonstrates the importance of presenting faculty members with guidelines and instruction when asking them to assess medical students. While this may seem obvious, and is widely considered good practice for assessment already, I think these findings provide insight into how faculty members will fill in the gaps with their own judgment of what's an appropriate way to assess students. This is relevant particularly for less-structured assessment opportunities, such as work-place based assessments in the real-world, clinical environment. Clear, complete criteria for assessment could reduce the number of gaps they have to fill on their own and improve the consistency and validity of their assessments.

Inference

The term "inference" is also borrowed from the work of Kogan et al. (2011), where she and others explored how faculty assessors assess residents. In both that work and my study, participants drew conclusions about the student's performance based on assumptions. While every participant made some sort of inference during the encounter, more-experienced faculty made them more frequently than those with less experience. Sometimes these inferences were explicit and were framed with statements, such as "it seems" or "I think" or "I'm assuming." For example, one participant stated "I assume this is [their] usual, approach and that hopefully that means that [the student] has like that step kind of down and organized and then [they]'ll be able to take it beyond that reporter stage" (Participant 8). Other participants did not seem aware that they were making such assumptions. For example, "The student's made a strategic decision to start, as students will, to start diving into different dimensions of the symptom without really, without really obtaining a sense of what the patient's story is or how the chronology of the history unfolds" (Participant 6). While the participant has made a judgment that the student has not yet collected enough information to move into "diving into different dimensions of the symptom," they are also assuming that the student made a strategic decision in choosing to move on, thus an inference.

Participants also commented on what they "hoped" or anticipated what the student would do in the encounter and seemingly made assumptions that these things would happen or were happening. One participant stated, "Hopefully, what's happening in his head is he is constructing potential diagnoses based on the history that he's heard..." (Participant 12). These inferences had a slightly different flavor in that they overwhelmingly seemed to operate in the student's favor and left room for the student to fill the gap, whereas the assumptions mentioned above were more often noting something negative.

Inference Summary

The *inference* theme exposes how clinical reasoning is an internal cognitive process, which makes it more difficult to assess than an observable behavior, such as doing a physical exam. In a patient encounter (simulated or real), we can catch glimpses of what the student is thinking through the questions they ask and what they explicitly state to the patient, however, this likely does not capture everything going on in the student's head. To fill in those gaps, assessors draw their own conclusions about why students conduct encounters the way they do, and these inferences can be both negative and positive. I believe that this becomes an issue for fair and valid assessment, however, when faculty are unaware that they are making these conclusions or when they base the majority of their assessment on assumptions as opposed to what they have heard or seen from the student. This will be discussed further in the i*ntegration* section below. It is also necessary for me to mention that throughout the analysis process, as with all qualitative work, I was also making inferences of my own about the study participants. These inferences color my own interpretation of the data and how I present it here.

Knowledge of clinical reasoning

Given the task for the study was to assess the medical student's clinical reasoning ability, it is no surprise that faculty drew on their own knowledge of what clinical reasoning is when assessing the student. Some participants used more academic clinical reasoning terminology, or terminology around clinical reasoning that would be encountered in an educational setting but perhaps not a clinical environment, when assessing student performance, while others did not. One participant commented:

> He got new data and he kind of broke out of whatever he was thinking and now is reevaluating his differential, which shows on some metacognition and his clinical reasoning skills. He's, he's double checking things and thinking about how that information changes what he's going to do and what he thinks the most likely diagnosis is. (Participant 14)

The use of the terms "metacognition" and description of the process of hypothetico-deductive reasoning (although that is not what he calls it) demonstrated an understanding of clinical reasoning beyond what would be expected of a clinician in a community-based setting. This is because while physicians in community-based settings learn how to *do* clinical reasoning, they are not all explicitly taught how the process works or is applied by learners and are more commonly using patter recognition as opposed to a more analytical approach. Other clinical reasoning terminology seen in the data include "illness scripts," "premature closure," and "hypothesis driven," all of which are commonly used terms in clinical reasoning teaching and research.

Daniel et al. (2020) identified seven steps in the clinical reasoning process: information gathering; hypothesis generation; problem representation; differential diagnosis; leading or working diagnosis; diagnostic justification; and management and treatment. All of these terms were used in a third pass of coding, however, only some of the terms or steps were found within the transcripts. Less-experienced participants were slightly more likely to refer to these steps than the more-experienced participants. However, when we look across all seven steps, at least one participant from the group of more-experienced faculty commented on six of the seven steps, whereas the less-experienced group collectively only commented on four. No participant mentioned step five, leading or working diagnosis.

I believe that the reason "differential diagnosis" was mentioned so frequently by participants (all but one) is that it is far more commonly used in regular clinical practice. None of the participants were presented with the Daniel et al. (2020) framework at the start of the encounter, so there was no expectation that participants would refer to these steps; however, as I try to understand how faculty are assessing the clinical reasoning process, the lack of their representation in the data makes me wonder if those steps can be measured via direct observation and how well they are understood by general faculty members within the School of Medicine.

Participants also used their own understanding of clinical reasoning or belief of what it is to help them determine what was relevant in their assessment and what to pay attention to. For example, while one participant commented that "In terms of, you know, clinical reasoning, it's important to get information in the right way, so [the student]'s off to a good start" (Participant 12). While another stated, "Not sure how much we should be kind of grading him on his clinical reasoning versus, like, just, you know, history taking and presentation skills" (Participant 9). The excerpt from Participant 12 demonstrates what they deem important and relevant to clinical reasoning, specifically highlighting the importance of gathering information. Conversely, Participant 9 implies that history taking (which is also information gathering) is separate or different from clinical reasoning. These differing views demonstrate the range of perspectives with which the participants came into the study.

Knowledge of Clinical Reasoning Summary

Knowledge of clinical reasoning, including the use of clinical reasoning terminology and the participant's opinion of what clinical reasoning is, came up fairly evenly across participants of different experience levels and all but one participant used language that was coded in this way. It is possible that the use of clinical reasoning terminology during the TA indicates that the participant has a better understanding of what clinical reasoning is, and therefore a better understanding of the construct they were asked to assess. However, because no formal assessment of academic clinical reasoning knowledge was provided to participants, I cannot know for sure. I also do not know if knowledge of clinical reasoning, which was inferred by the use of academic language, allowed participants to provide a "better" assessment of the student's clinical reasoning ability, which is subjective and will be discussed further when considering the theme "judgment."

Assessor Factors Summary

The *assessor factors* theme highlights the differences across individuals by encompassing the internal components each individual brings to an assessment opportunity that are unique to only them. These components reflect years of experience, collected knowledge, and ways of thinking that cannot be altered by external factors, but potentially can be managed. For example, if I were asking a group of faculty to conduct an assessment, I cannot know the exact level of clinical reasoning knowledge each of them come into the room with. I can, however, provide them with a baseline level of understanding of what the construct is, as well as what it is not. I can also provide frames of reference that calibrate each individual assessors opinion of how they *think* the student should perform to instead a standard, consistent expectation. Lastly, I can ask them to be aware of the inferences they are making, and guide them to focus on what they hear and see in the encounter.

Because the assessor is the one making the ultimate decision about the student's performance, it's challenging to separate the categories noted above within *assessor factors* from how faculty are processing what they are seeing (*integration)* to ultimately assess the student. For this reason, before moving into the other factors I will discuss the *integration* theme in the next section.

Integration

I have defined the *integration* theme as the decision-making process participants engage in when assessing a student before making a judgment about their performance. This theme captures how faculty are interpreting the information they are receiving from the student and situation (*student factors and situational factors*) with their own experience and knowledge (*assessor factors*) to ultimately make a judgment. When participants engaged in *integration* they appeared to either be incorporating multiple pieces of information together or trying to make sense of conflicting information they were receiving from the student's actions.

Participants took different approaches in the reasoning process. Some noted whether the student arrived at the "right" diagnosis, for example, "I think he ended up nailing it in the end in regards to the differential" (Participant 7). Others focused more on the cognitive process the student used to arrive at their conclusions. A participant noted, "It seems like the thing that is in my mind not quite adding up as he seems to be focused on the appendix and ruling out the appendix when he has a lot of information in front of him that it is not the appendix" (Participant 5). Another stated,

> [Student asked] that sort of pseudo question about diarrhea, but that should have been part of the main HPI. But he has it embedded in review of systems which they started with swallowing problems. So I don't know if he's having a hard time deciding if he's in a GI inquiry like gastrointestinal inquiry vs.in the explicit review of systems section of a standard history so not clear, I would actually ask him about that if I were doing an actual debrief. (Participant 4)

Of the participants that did comment on whether or not the student got the right diagnosis, only one did not explicitly comment on how they were interpreting the student's reasoning process throughout out the encounter. All others expressed concern about whether or not the student would arrive at the right conclusion, but also commented on how they were considering other facets of the encounter (i.e., student's choice of questions). Ultimately, however, the student stating the right diagnosis at the end of the encounter seemed to weigh heavily. For example, the same participant from above, Participant 4, also commented, "I guess we'll see in the final and the final accounting what he picks, but it seems like he's going to pick PID" (Participant 4).

As the faculty grappled with how to assess the student's clinical reasoning process, they sometimes used words such as "it's interesting" or "I'm curious" before sharing their thoughts. I interpreted these as cues that they had picked up on something important and were digesting what they had just heard or saw. Participants, particularly those in the more-experienced group, also seemed to struggle with their own uncertainty during the encounter. While it seems that in some points of the encounter they felt comfortable making inferences about the student's choices (see the *assessor factors* section above), at others they expressed wanting more information or being unclear on how to interpret what the student did. One participant stated:

> He's like, fair at gathering information, like he got the right information…but kind of at a novice level, like the information was obvious and that's why he got it, but I wonder that if, if the signs were more subtle, if he would have necessarily came to the right conclusion. (Participant 13)

Several participants mentioned that, if possible, they would have liked to ask the student more questions to aid in their assessment and reduce some of their uncertainty.

Another type of uncertainty was around what was "fair" to assess the student on and what really counted as clinical reasoning. In addition to the concern noted above about specialty differences in expectations, one participant noted, "I'm guessing because he's like more of a medical student level that we're not really judging them too much on the true like clinical reasoning or clinical decision making" (Participant 9). When participants were waiting on certain pieces of information, and they finally received it from the encounter, they would often express agreement with the student, "But I think after all the questioning, he's arrived at the right preliminary ideas" (Participant 1) and seemed much more comfortable in their decision making. For example, after hearing a statement from the student, one participant said:

> So for me that I mean, basically, we could have skipped a lot of the early part and just got to that 30 seconds like that that kind of shows you everything. So he's managed to narrow it down to kind of two big things on his differential at this point. (Participant 14)

Integration Summary

Each participant's integration or decision-making process varied. The prompt of the interview was for participants to assess the student's clinical reasoning ability. Participants interpreted this to either be whether or not the student got the "right" answer or how the student got to the "right" answer. Although I am not surprised that some faculty focused on the student's arrival at the "right" diagnosis, which interestingly in this situation was an unknown for the faculty member and there was no way for them to verify or confirm it, I argue that using that approach is not truly assessing the clinical reasoning *process*. For those who did try to assess the student's process, which was most of the participants, I found the method they used to be very similar to the hypothetico-deductive model of clinical reasoning, where individuals make a hypothesis about a diagnosis based on provided information and then collect more information to either support or refute their hypothesis. This process is iterative until they are able to hone in on their differential based on the data available. Similarly, participants seemed to make some early estimates of how the student was doing based on initial lines of questioning and then observed the student's subsequent questioning, sometimes seeking out specific information to validate their estimate or anything that would confirm or adjust their initial assessment. This happened throughout the encounter until the end when they would reach some sort of final judgment, which will be discussed in the *judgment* theme.

Assessor and Integration Summary

To me, one of the most surprising findings of this study is the variability in approach and thinking across all the participants. As a non-physician, I expected to see a lot more consistency across the ways in which the participants would assess the student. I believe this expectation came from the belief that there was one right way to collect a history or diagnose a patient. However, examining the data demonstrates the unique approach each individual takes to not only assessing the student, but how they think about making diagnoses. The *assessor factors* noted above that each participant brought with them into the scenario color their interpretation of what the student is doing and how they should be assessed. Therefore, we can expect variability in how faculty members integrate all the information they are receiving to make an assessment judgment because each faculty member is unique. In the next section, I explore factors external to the individual assessor that the faculty observe and react to as part of the integration process, *student factors* and *situational factors.*

Student Factors

The theme *student factors* captures the elements that the student, the object of the assessment, contributed to the participant's assessment decision. This theme included three categories: student organization, student communication, and student-SP interaction. All participants' transcripts contained language relating to the codes within this category.

As noted above, while these factors were driven by the students actions and comments, because the assessor or faculty member is interpreting everything the student does or says through their own lens, these factors interact with the themes mentioned above, *assessor factors* and *integration*. Unlike *assessor factors,* however, these factors (along with *situational factors)* are external to the individual completing the assessment.

Student Organization

Student organization includes how the students structured the encounter with the patient in terms of how they collected information and transitioned through different stages of the encounter (i.e. history of present illness (HPI) into review of systems (ROS) into physical exam (PE) into counseling, etc.) Participants commented frequently on how the student directed the encounter through the order in which they asked questions and at what phases of the encounter they collected certain pieces of information. One participant noted, “So because you think of the review of systems as really just like a list of anything else that you didn't catch. So I think the components that he mentioned about for ob/gyn history would be maybe something better to put in the HPI specifically” (Participant 10). They also frequently commented on how the student organized their questioning and how their organization scheme either helped ensure the student would get the right information or their lack of organization would prevent them from capturing information they needed. One example includes:

> So he got there, but his organization is really is challenging…He, he had to talk it through the action get his thoughts organized enough. And I think one of those challenges is that he didn't have any sort of organization to help him. (Participant 3)

All but two participants noted when the student did not collect information from the history that they felt was critical to the case. One participant stated, “He didn't quite hit every single question I wanted him to hit” (Participant 2).

Participants also frequently commented on the use of algorithms or templates. Most often the student’s use of the onset-provokes-quality-radiates-seveirty-time-alleviating-aggravating/ alleviating-associated-attributions (OPQRSTAAA) checklist for pain assessment was noted positively. Participant 9 commented, “So I'm just picking up the he's kind of going through the appropriate questions for pain assessment. So that's good. It seems like he's being thorough and trying to hit all the major points there” (Participant 9). While the use of templates was frequently noted as being a positive because it offered more organization and structure for the student, faculty also appreciated when the student veered from the “typical” checklist in a focused way:

> He sort of sees where she's going so and I think from a reasoning standpoint and seeing him start to be able to veer from a script is usually a good thing. If he’s veering off in a little bit of a focused way. (Participant 8)

This student’s ability to focus their inquiry was something some participants seemed to be watching for. For example, one participant noted, “I think he's just not, not thinking particularly about a differential, but he's trying to gather all the basics…and I'm hoping that after that he can zoom in and focus on pertinent questions more thinking about a differential” (Participant 1). Interestingly, using a template or being too scripted was also noted as a negative by some participants. This was linked to the use of close-ended, checklist-like questioning that will be discussed further below.

Participants seemed to use the way in which the student organized their questioning and moved through the encounter as a representation of how the student had organized information about the patient. When the student moved seamlessly from one area of questioning to another with no missing information, participants noted the student to be organized and “spot on” (Participant 7). However, if the student missed questions or the flow of questioning seemed disjointed, participants noted the student’s organization to be “challenging” (Participant 3).

Student Organization Summary

Prior to conducting this study, I would have hypothesized that student organization would have had the most consistent findings across all the participants. I would have argued that this is the area with the most objectivity because students are taught to ask questions and organize the encounter in a specific way and have been trained to tailor their questioning according to the information they receive from the patient, which is universal across medicine. I was, however, surprised to find so much variability in the participants' reactions to the student's approach. The most commonly agreed upon "good" thing that the student did was that they followed the checklist for gathering information about pain, but even that was not unanimous and some participants felt like the student could have collected better, more complete information had they "let the patient tell her story" (Participant 6) and not asked things so directly. One could argue that the lack of checklist contributed to the lack objectivity, however, I would argue that students and the faculty who assess them do not receive guidelines or a checklist for how to diagnose every patient they encounter in the hospital or clinic, so the findings are still relevant and important. As students bounce between different preceptors, the criteria that they will be assessed against will change and will likely not be made explicit.

Student Communication and Student-SP Interaction

All of the participants had some observation about how the student interacted with the standardized patient, the language they used, and the student's general comfort level during the encounter. Because these categories are so closely intertwined, for the purposes of this discussion I have combined the Student Communication and Student-SP Interaction categories and will discuss them together.

Many participants stated that the student looked "awkward," "nervous," or "uncomfortable." Several faculty also noted that the student took several long pauses, with one commenting, “This is a very long pause for him. He's like trying to gather his thoughts or figure out what to do next. But that was a pretty long pause” (Participant 8). Interestingly another participant noted at the same moment in the encounter that it was “good that [the student]’s taking the time” (Participant 3).

Some faculty honed in on the language the student used in the encounter. This included how the student referred to the patient (i.e., “girl” instead of woman or female), but also general phrasing as he shared information with the patient. The most commonly noted concern around communication was the use of open- vs. closed-ended questions. One participant stated:

> As he's talking I'm already paying attention to, like, how many closed ended questions he's asked. There seemed to be a lot more, he started saying like will tell me more about that, but then quickly transitioned to closed-ended questions and I'm sort of paying attention to, to, it seems like he is prompting her a lot with like a, like, oh like down by the hip area or sort of like an aching pain. (Participant 13)

Similarly, others noted the use of leading questions when the patient would ask questions directly, such as “do you get a rash [when you take penicillin]” as opposed to asking “what happens when you take penicillin?” Participant 5 noted, “He's having a little bit of leading questions and he seems a little awkward.”

Participants also commented when the student stated something in a way that they thought was unusual, awkward, or strange or used a lot of fillers. One attributed the student’s language, as mentioned above, to being a key driver of the awkwardness, but was less critical because of the (assumed) level of the learner. The participant stated, “He has a lot of hesitation. A lot of ‘ums’, and ‘kind ofs’ almost creating a little bit of an awkward scenario at times, but I know again that just comes with time” (Participant 7). Language was also linked to the lack of confidence the student was portraying to the patient. One participant suggested the student "take a second to get [their] thoughts in order before speaking, i.e. 'some people get ovarian abscesses' sounded awkward, 'umm'/'like'/ 'I feel like we can help you' do not build confidence for the patient" (Participant 2).

Participants also commented more generally on how the student was portraying confidence to the (standardized) patient and how this influenced his ability to establish a positive rapport. One participant noted,

> I think he's just like, really uncomfortable asking the questions that need to get asked. And so he's like saying things that it's sort of a stating the obvious in a way that I think kind of undermines his own credibility. (Participant 11)

More positively, several participants did note when the student capitalized on opportunities to enhance his relationship with the patient. Participant 6 stated, "I think [the student] created a nice space in the initial moments for this patient to tell her story." However, several, also commented when the student missed an opportunity to create rapport. For example, one participant stated, "[The student] hasn't expressed any sort of empathy around the fact that she said that she had terrible stomach pains" (Participant 3).

Student Communication and Student-SP Summary

A student's ability to positively interact with their patient is a fundamental skill that is assessed in the earliest stages of medical school. Before students have acquired the biomedical knowledge they need to diagnose, students first learn how to ask patients questions and gather information in an organized manner. When we examine how this relates to the assessment of clinical reasoning, I would argue that the first stage of clinical reasoning, information gathering (Daniel et al., 2019), is directly related to how a student goes about collecting data and that their approach to doing so (i.e., the way they ask questions, the rapport they develop with the patient) can impact the quality of data collected. Some participants explicitly made this connection, stating, “I think when it comes to clinical reasoning you're less likely to get a good history out of a patient who doesn't feel like you care” (Participant 3), while others noted the opposite, “I would have phrased that differently is what I'm thinking, as far as he kind of led the patient at that point. But really doesn't have anything to do with clinical reasoning” (Participant 5). Others made a number of comments about the student’s use of language, but never explicitly stated if they thought it was or was not relevant to the task at hand, the assessment of the student’s clinical reasoning ability. For these participants, I do wonder if they may have been assessing a different, but related, construct altogether, perhaps communication.

Student Factors Summary

Although not specifically relevant to clinical reasoning, student communication and student-SP interaction was noted by all but one participant. While I can only speculate as to why communication and student-SP interaction seemed to be such an area of focus, I wonder if an increasing emphasis on patient-doctor, or patient-student, communication in medical training over the last five to ten years may contribute to this finding. Additionally, while all three are student factors, student organization seems to be more directly correlated to a tangible skill that a student would have been taught and expected to present. For example, when first learning how to take a history, students are provided with a template of what questions to ask and in what order. Over time students are expected to rely on this template less and less and tailor their questioning to the patient in front of them, but there is a backbone as to the process they use. Student communication and interaction with the patient, however, are much more subjective, influenced by the student’s own personality, and likely assessed based on personal opinion.

As the subjects of any exam, students of course influence the outcome of assessment. Examining student factors within this study demonstrates that beyond knowledge and skills, faculty are also interpreting other components unique to the individual. Some of these components, communication and interaction, are much more subjective; others (the student's use of questions, flow through the encounter, and adherence to a template) are more objective. I believe that helping faculty members become more aware of these subjective components can help them shift away from focusing on what could be argued to be personal preference, or at least acknowledge the influence of those preferences on their assessment of the student.

Situational Factors

The next theme I identified is *situational factors*. These factors are not necessarily tied to either the assessor (faculty participant) or assessee (student) but to the specific scenario or setting in which the assessment is taking place. I identified four categories of situational factors: context, cognitive load, specialty influence, and limitations of the study. Context refers to the setting or location in which the case is supposed to take place (i.e. emergency room vs inpatient ward vs outpatient clinic). Cognitive load describes the competing mental demands on the assessor during the task. Specialty influence captures the ways the specialty in which the participant trained influenced their approach in the encounter. Lastly, limitations of the study shows how the study design itself affected the decision-making of the assessor. The codes identified within this theme were not as apparent as other themes previously discussed. They did, however, feel important to include because when they were present they seemed to influence the participant's assessment.

Similar to *student factors*, *situational factors* rely external to the assessor, but still greatly influence their decision making process. While they will be distinct to each assessment scenario or the structure of the case, they still interact with how the assessor integrates what they are seeing and therefore influence the faculty reasoning process.

Context

The context in which the case was set (i.e., inpatient, outpatient, emergent) factored into the decision-making of several participants. In the door note provided at the start of the encounter, participants were told that this encounter was supposed to be treated as if the patient and student were in the emergency room. This information appeared to have no effect on some participants, while it factored greatly into others' assessment decisions. One participant, for example, was so concerned with the student not tailoring their approach to what the faculty felt was a serious, emergent issue that it colored their entire perspective of the student's performance. This faculty member commented:

> I'm not getting that like when you have potentially emergent conditions, when you have time sensitive conditions, there are questions that are kind of a waste of time, right? Like if he starts asking about her family history I'm gonna blow a fuse. (Participant 11)

Another faculty member failed to note, or perhaps forgot, that the encounter was supposed to take place in the emergency room and gave the student credit towards their clinical reasoning for doing a pelvic exam, which is "clearly not an exam that you do standard in your office. So I do think he was he was hypothesis driven" (Participant 8), implying that they felt the exam was supposed to be set in a traditional outpatient clinic. A couple other participants asked for clarification on where the case was supposed to take place later on in the encounter, which I interpreted as they forgot that information had already been shared, but also that something triggered them to consider that that piece of information was relevant for their assessment.

Context Summary

The influence of the context of the case on the participant's assessment of the student varied. As noted above, one participant greatly weighted their assessment of the student because of the context, while others seemed to not even take into account. In a real-life scenario, this of course would not be an issue, however, it does make me wonder if students are prepared for adjusting their reasoning process and approach to fit the different spaces they will find themselves in. The comments that came from the data when discussing context also imply that there are some things that are okay for the student to miss, and others are not. Explicitly noting what those things are is important for student understanding, but also for faculty to be aware of as they assess an encounter.

Cognitive Load

Cognitive load describes the competing mental demands on an individual when completing one more complex tasks. This overwhelm can lead to misremembering or the inability to recall information. Within the encounter, numerous faculty commented on questions the student missed or did not ask, despite the student having asked them and the patient responding. In these cases the faculty's misses were attributed to the student's performance. For example, despite the student asking about surgical histories and receiving an answer form the patient, one participant stated, "I don't know if he asked about passed surgical histories or maybe he'll ask about that later. So that was something that I didn't quite catch" (Participant 10). In this example, the participant seems to attribute the miss to themselves, rather than the student. However, in other cases the participant believed the student just missed the question, "So even though he did kind of march down his like checklist I don't think he asked if she's ever had any surgeries in the past, right?" (Participant 11).

Cognitive Load Summary

Participants in this study were not given any additional information about the case prior to receiving the same door note that the student received at the start of the encounter. As a result, they had to simultaneously diagnose the patient *and* assess the student's clinical reasoning ability. This extra demand on their thinking could be due to overtaxing their cognitive load. Cognitive load theory argues that individuals can only process a finite number of elements or information at any given time (J. Q. Young et al., 2014). One participant even commented that they felt it was easier to assess students when they had a checklist in hand or knew about the case beforehand. However, some argue that using checklists can actually increase an assessor's cognitive load rather than diminish it (Gingerich et al., 2014). I argue that the structure of this study forced faculty members to assess the student more similarly to how they would in a real-life patient encounter as opposed to a standardized exam, and therefore reflects the cognitive demand that is placed on them when they assess students in the clinical setting. This could have significant implications for assessment validity.

Specialty Influence

I intentionally recruited participants across a number of specialties to see how specialty training may influence faculty decision-making around clinical reasoning assessment. The focus of the case was lower abdominal pain, but all participants quickly realized the case required knowledge of women's health. Interestingly, the subject matter of the case (women's health) influenced the participants comfort level with assessing the student in different ways. One participant who sees women's health issues regularly in their practice noted that they were going to be "way pickier on the pelvic stuff" but also felt that "might be unfair" (Participant 2) to assess them against. Another faculty member felt that they needed to qualify their assessment because they are “never seeing this type of presentation in [their] clinic” (Participant 7), while another felt like their specialty might be making them not be critical enough, stating, “I can't off the top of my head think of any major things he didn’t ask. Although I’m certain an internist or surgeon could” (Participant 12).

Specialty Influence Summary

In the real clinical setting, faculty often precept medical students in the specialty of their choice. For standardized exams, however, medical schools often focus recruitment on faculty who have experience in medical education and assessment and their specialty is secondary, if at all considered. The cases they are asked to assess can also vary widely in their content (i.e., chronic disease management vs acute illness, general concerns vs women’s health, etc.). While the assessors are *sometimes* provided with faculty development prior to the case so they can prepare, to my knowledge the preparation does not account for calibration of specialty specific knowledge that might influence their interpretation of the student’s performance.

Limitations of the Study

Although small, I felt it was important to note that the format of the encounter (a recording of a video encounter via Zoom) felt limiting to some of the participants. Without being able to see the patient fully and have a good angle on the student, they felt they could not make a good, fair assessment of their clinical reasoning ability. Because the recording was also with a standardized patient, one faculty noted that the student may be behaving differently than they might with a real patient because students might be doing what they think they are “supposed to do” for the assessment (Participant 14). Few participants noted these limitations, and of those who did, many did not mention it again once they got deeper into the encounter. One or two participants commented on this limitation towards the end, but it was more in the context of wanting to be able to discuss the student's performance directly with the student before providing feedback.

Limitations of the Study Summary

While noted far less frequently in the data than the previously identified factors, situational factors seemed to have a significant influence on participants when it came up. As these factors are outside the faculty member's control, the educators designing assessments should be mindful of how they influence the faculty reasoning process and what they can do to mitigate unequal influence across assessors.

Situational Factors Summary

The *situational factors* theme had the most inconsistent prevalence across the data. However, as noted for the context and cognitive load categories, when present I would argue they significantly influenced the assessor's integration and decision-making process. I believe that the categories noted within this theme also reflect potentially the easiest opportunities for influencing assessment encounters, since they do not require changing behaviors or beliefs of individuals. This will be discussed further in the following chapter.

Judgment

The next theme I identified was *judgment*, which I noted as anytime a participant made qualitative statement about the student's performance. Unlike the *integration* theme described earlier in the chapter, and excerpts from this theme capture dhow participants were grappling with information and what to do with it, *judgment* represents an actual decision or conclusion to the integration process that was then shared in the TA interview. Two types of *judgments* were noted within the data, micro-judgments and summative judgments. I identified micro-judgments as the small, in-the-moment assessment decisions participants made about the student's performance throughout the encounter. Conversely, summative judgments were final decisions or statements about the student's performance closer to the end of the interview. These judgments were coded as either positive, negative or neutral and almost all participants made both positive or negative summative or micro-judgments throughout the encounter.

Micro-judgments

Half of the participants made both positive and negative micro-judgments of the student's performance at varying points within the encounter. One example of a negative micro-judgment made by a participant is:

> So whenever there's a medical problem and medical diagnosis, what I like to hear from students or residents is the why question, try to understand why this patient has what they have. And I don't think he's digging deep enough to understand why she has this…and that is crucial. (Participant 1)

An example of a positive micro-judgment that occurred early on in the encounter is "So at this point I'm from a clinical reasoning standpoint providing him with higher scores, because I do feel like his questions are relevant" (Participant 7). Of the faculty members who did make micro-judgments throughout the encounter, they often remained positive or remained negative throughout the encounter. For example, only two of the participants made both positive and negative micro-judgments, two made only a single micro-judgment (one positive and one negative) , while the others made multiple only positive (three participants) or only negative (two participants) micro-judgments. Interestingly, Participant 11 was the only participant to make consistently negative micro-judgments and state a negative summative judgment, which will be discussed further in the next section.

Summative judgments

Participants were not directly asked to rate the student's clinical reasoning ability on a scale or provide a summative judgment statement. However, some participants, because the task was to assess clinical reasoning, did provide statements about the student's performance overall. These summative judgments varied from positive ("excellent" or "satisfactory") to negative ("failing"), with some mixed ratings in between. For example, when reflecting on the student's performance, one participant stated, "So in the grand scheme of things, this isn't a fail and [I] certainly have had excellent students who became excellent residents who didn't do great on this case" (Participant 4). Of the participants who explicitly made summative judgment statements, all but two were positive. Of the remaining two, one was negative and one was neutral. As noted above, the participant making the negative summative judgment (Participant 11) made several negative micro-judgments throughout the encounter. Participant 13 made a neutral statement, however, and made no micro-judgments throughout the transcript.

Judgment Summary

Similar to the unexpected variability in integration processes, the variety in judgments, both micro- and summative, is possibly the most striking finding from my data. Before engaging in this study, I would have guessed that overwhelmingly the participants would arrive at similar conclusions about the student's performance, particularly since this was a standardized case created with the intent of assessing a novice-intermediate learner. While the majority of the participants arrived at a positive summative judgment of the student's performance, the range of adjectives used to describe the student's overall performance (noted above) indicated that there was significant variability across the fourteen participants. In addition, examining the micro-judgments allowed me to identify moments where participants would interpret the same statement or question differently. As a non-physician this is very surprising. While I would expect the clinical reasoning process to be more subjective when influenced by *situational factors* within the real patient-care setting, I expected it to be much more objective for the case presented within the video.

The micro-judgments made by the participants also indicate the potential for confirmation bias or influence of the halo effect on assessors. In clinical reasoning, confirmation bias is "the tendency to look for evidence to confirm a diagnostic hypothesis rather than evidence to refute it" (Thampy et al., 2019, p. 1632). In assessment, this occurs when faculty members make an initial judgment about the student and then look to confirm the initial judgment for the remainder of the assessment. For those how made initial assessments, or micro-judgments, once made, participants seemed to be evolving their judgment on an imaginary scale from "bad" to "good." For example, as seen in Table 4, the two participants who made both positive and negative statements started out with positive micro-judgments, then made negative micro-judgments, before ultimately making differing summative judgments.

Table 4. Evolution of judgment statements from participants 1 and 14

Judgment	**Participant 1**	**Participant 14**
Positive Micro-judgment	"It seems like he's in the middle of gathering all the pertinent information And so far, he's doing a good job."	"He's also asking about things related to STIs because that's on the differential and so I'm yeah I'm seeing pretty good stuff here in terms of clinical reasoning."
Negative Micro-judgment	"And here, he's not deciding to investigate that any further, which I think is not a very good idea"	"He's changing his workup from what he told the patient about… This is just one of those like what happened…it makes me want to rate him a little lower"
Positive Summative Judgment	"I think he came to the right conclusion of what she probably has but he's not answering the why question and that is crucial."	"Overall I would say you have good to excellent clinical reasoning skills for a medical student and are ready to progress to intern year."

Depending on that initial statement, the starting point on their internal scale, may influence how likely they are to adjust their criteria enough (if warranted) for a student to go from good to bad or vice versa. This anchoring effect, or “the tendency to over rely on, and base decisions on, the first piece of information elicited/offered” (Thampy et al., 2019, p. 1632), appears to be stronger when the anchors are self-generated (Epley & Gilovich, 2001). It is possible that participants would have been affected by anchoring since they were not provided with any scales or anchors during the assessment task.

Feedback

The category of codes identified within the data related to the type and quality of narrative feedback participants provided at the end of the encounter. The only guidance they received was to treat the feedback form (in Qualtrics) similar to how they would provide feedback to students in the electronic evaluation system (MedHub). Within the written comments, participants used different feedback styles and tones, which were coded separately. The most notable differences were the use of direct vs indirect statements and active vs passive constructive feedback.

Participants were split fairly evenly in whether they used direct feedback (i.e., “you”) or indirect feedback (i.e., “the student”, “they”), while one participant switched mid-way from indirect to direct. Because the participants had no direct contact with the student, I think the use of indirect feedback is not surprising. This style is, however, very different than feedback that is traditionally provided on real clinical rotations or assessments. There was also no pattern within the data relating positive or negative judgments to indirect or direct feedback. I could imagine that participants with less good feedback to provide, may choose to use a more indirect stance, but that was not the case in the data for this study.

I also noted a difference in tone when participants provided constructive feedback. Some faculty chose to use softer, more passive language, for example, “I would've liked you to explore more gyn history earlier, as well as full treatment history of gonorrhea infection” (Participant 13). Others used stronger, more active language, such as, “You asked about sexual history and menstrual history, but didn't ask about contraception or history of pregnancy” (Participant 12) clearly indicating a missed piece of data collection. Interestingly, participants that used direct language for the feedback were more likely to use an active tone when providing corrections or suggestions for improvement compared to those who chose to use an indirect style in their narrative comments.

Some participants also commented on general principles that they have when providing feedback. For example, a couple participants commented on the fact that they would usually provide the student with feedback face-to-face before putting anything in writing. Participant 14 stated, “So if I don't have a chance to talk with them, I don't know what to do with the, you know, still continuing to worry about appendicitis quite so much and I always struggle with what to put in an evaluation form about something like that with the information that I have at this point, um, because I'm basically going to be making conclusions without enough information.” Participants acknowledged that this was just a limitation of the study, although I thought it was interesting that they felt they had to vocalize this concern when the student would never receive the feedback they were providing.

Feedback Summary

While understanding how faculty members frame their assessment for students was not a direct aim of this study, it provides insight into how the decision-making process and judgments are translated to students. While in-person feedback is highly encouraged, it does not always occur, leaving the student with only a checklist and/or some narrative comments, depending on the situation. The patterns within the data also provide important implications for practice. The most notable implication is the opportunity to create more consistency in the type of language and tone used across assessors. This is further discussed in the following chapter.

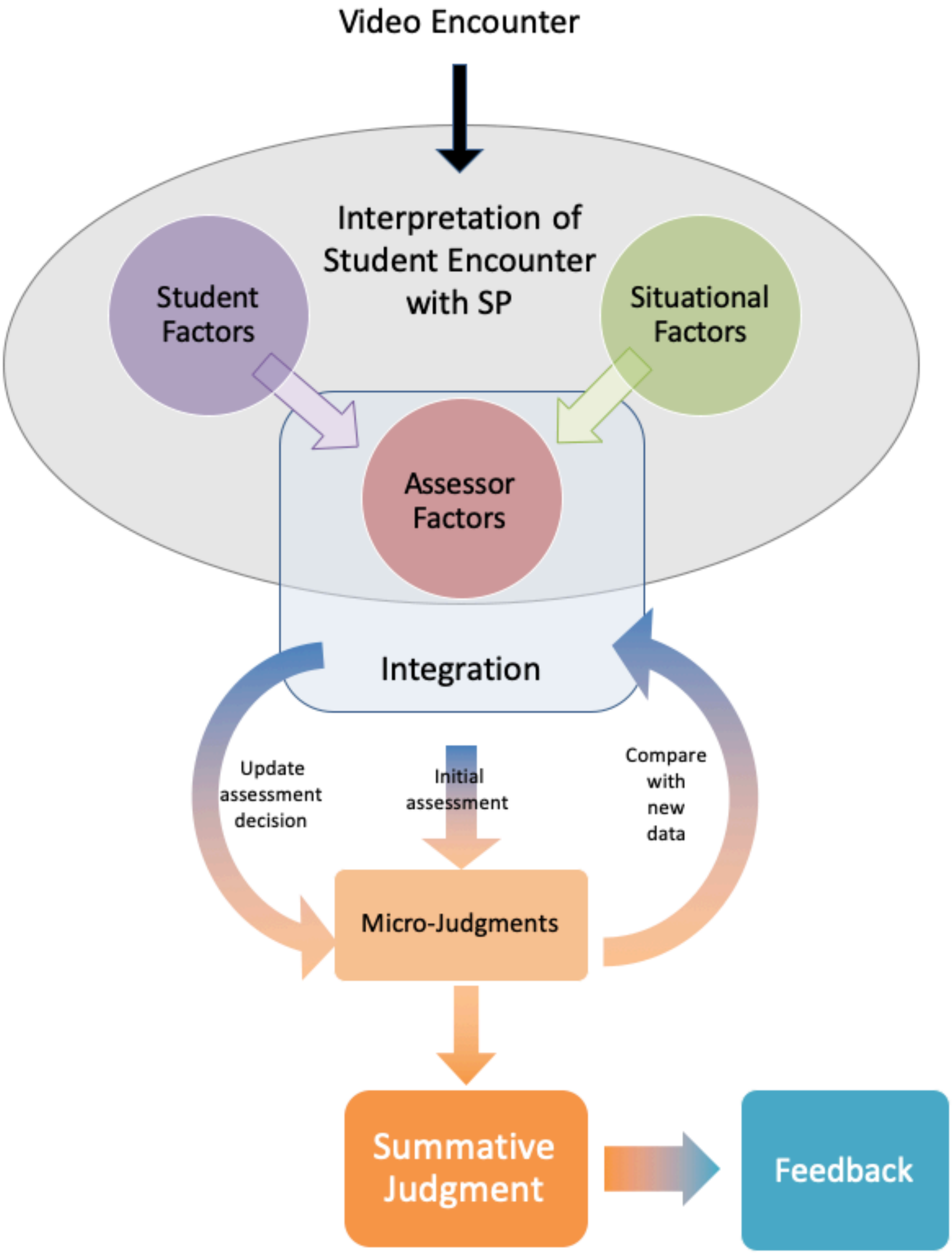
Now that each theme has been defined and broken down, I will use the next chapter to discuss how the themes interact with one another in what I have called the faculty reasoning process. In addition, I will explore the implications of my findings on current assessment practices within medical education and how each of the themes may potentially be influenced to mitigate bias and increase validity.

V. The Faculty Reasoning Process and Implications for Practice

The aim of this study was to better understand how faculty assessors make assessment decisions about medical student's clinical reasoning ability, particularly in the context of clinical encounters. In addition, I hoped to understand commonalities and differences across different participants in hopes of understanding where variation exists among different faculty members. By identifying these differences, we can explore opportunities for better standardization of how students are assessed which will hopefully lead to increased reliability and validity. I also hoped to identify any opportunities where lessons learned from this study may be applicable or opportunities to explore when thinking about assessment practices in the real-world setting (i.e., clinical encounters with real patients). The themes (which will be noted in this chapter in *italics)* I identified provide a better understanding of what factors influence that assessment decision-making process, what I call the faculty reasoning process (Figure 3). In the following section, I will walk through the model and identify opportunities for influencing or mitigating the effects of these different components. When examining how the various themes interact, it is helpful to examine them in groups.

Figure 3

Conceptual model of the Faculty Reasoning process



Factors and Integration

As described in the previous chapter, the multiple different factors come from different components of the encounter, either the assessor, the student, or the situation in which the case/encounter exists. While the *student factors* and *situational factors* originate outside of the individual assessor, the *assessor factors* cannot be separated from the faculty member conducting the assessment nor the process the use to interpret these factors and make an assessment decision. This process is *integration.* These factors and the process by which they are integrated significantly influence the faculty reasoning process.

No assessment is free from bias or influence, either by the individual who has created the assessment or the individual implementing the assessment. Each of us brings unique knowledge, experiences, and viewpoints that can easily color our perception and make us interpret what we see differently. Faculty assessors are no different. In the assessment of clinical reasoning, faculty bring in their own clinical reasoning process, which is influenced by their own training and experience, as well as their knowledge of clinical reasoning. They also tend to want to have some frame of reference, whether that be other students or themselves, and rely on inferences to interpret student behavior when lacking explicit knowledge of what the student is thinking. These factors are unique to the individual and cannot necessarily be controlled for when planning for an assessment or selecting assessors. By understanding these factors we can, however, moderate their influence.

One suggestion for moderating differences in clinical reasoning and knowledge of clinical reasoning is to offer focused faculty development on what clinical reasoning is and expectations for what reasoning process(es) is expected of the student. For example, because advanced physicians tend to use different approaches to clinical reasoning than novice learners due to their level of experience (i.e., pattern recognition vs hypothetico-deductive reasoning), informing faculty of the approach students are likely to use will allow them to hone in on how students are following that approach or pattern. Rubrics identifying the different phases of clinical reasoning may also be useful, as they can provide explicit guide points for faculty to reference throughout the encounter. Keeping these rubrics simple, however, is important to reduce the potential for added cognitive load on assessors (Gingerich et al., 2014). Being explicit about what counts in the assessment is also critical. While some OSCEs appropriately assess core communication and professionalism skills, it may be easy for faculty to focus in on those skills over the clinical reasoning process without explicit direction to do otherwise.

Situational factors play a smaller but still significant influence on the faculty reasoning process. These factors are unique to the context in which the case takes place, the biomedical and clinical topics associated with the case, and other demands competing for the assessor's attention. While assessor factors are based within the assessor, situational factors are based outside the individual and can interact with the other factors as well as the reasoning process. For example, within the study I noted significant interaction between the content of the case and specialty of the faculty member. This was perhaps more noticeable because the case was related to women's health. Had the case been centered around a more generic, bread-and-butter topic (i.e., cold or cough), perhaps this influence would have been less noticeable. As physicians move from medical school to residency to practice or fellowship, their training becomes more and more focused. While the case presented in this encounter may seem like it should be easy if a medical student can do it, some faculty may have not diagnosed a women's health issue in years. For those faculty members who were less comfortable with women's health, there may have been an increased demand on their cognitive load as they tried to reason through an unfamiliar case as well as assess a medical student.

In a true clinical setting, the influence of situational factors on the faculty reasoning process may actually be smaller than in a standardized OSCE, where the answer to the case is often given to assessors beforehand. Faculty members rarely precept and assess medical students in clinics or inpatient settings that are not within their own specialty/department, therefore the risk of them being presented with a patient with symptoms that are less familiar to them is greatly reduced. One could posit that faculty and students may still work with patients who present with less common symptoms or rarer conditions, in which case this could impact how they assess their students, but this is much less common. However, in my experience with administering standardized assessments, such as OSCEs, at the medical school level, particularly in the pre-clinical years, faculty are recruited from a broad range of specialties with the primary focus being on availability. While logistical and resource barriers prohibit soliciting assessors with case-specific knowledge, two ways to potentially mitigate the impact of situational factors on the assessment process are to orient faculty to the case and provide a minimum level of knowledge around the subject matter. In addition to filling in the gaps of their own knowledge, this would serve as a reminder of what knowledge the student would be expected to bring into the case. This could help decrease cognitive load by reducing the amount assessors have to pull information from their long-term memory while also trying to assess the student.

A reliable assessment, according to generalizability theory, limits the variability of other factors so the main source of variance is the object of assessment, or the student (Bandalos, 2018). In the context of the assessment used in this study, the student should be the primary component shaping the faculty member’s assessment decision. While the other factors certainly influence the faculty reasoning process, they are all, in theory, a reaction to what the student says or does. However, with the focus of this assessment being on the clinical reasoning process, I argue that other student-related factors, beyond clinical reasoning ability, can influence the assessor. Some of the faculty were distracted by the language the student used, pauses within the encounter, and lack of empathy expressed to the patient. There is no debate that these are important for the doctor-patient relationship, but are they important for clinical reasoning? Leaving this for assessors to decide for themselves can lead to significant variability and potential sources of construct irrelevance (Bandalos, 2018), impacting the validity of the exam.

Integration and judgment

Integration and *judgment* are at the center of the framework. Within *integration*, the factors described above are synthesized and merged with the already existing knowledge, opinions, and experience of the faculty assessor. Once the assessor has grappled with how to interpret this information, they make a decision about the student which comes out as a *judgment*. The cyclical arrows around *integration* and *judgment*, indicate that this can be an iterative process within the assessor's mind, a process that is similar to the hypothetico-deductive process of clinical reasoning.

With some amount of preliminary information, compared against a preconceived notion of what the student should do, assessors make an initial estimate of the student's ability, or a micro-judgment. This preconceived notion can be grounded in what they believe they themselves would do in the situation, or it can be based on what they would expect the student to do based on how other students have behaved in similar situation. They then observe the student's behaviors and questions and compare that to their initial assessment and adjust accordingly. This series of reasoning and judgments ultimately leads them to their summative assessment decision.

Also similar to the clinical reasoning process, some assessors are simultaneously evaluating the quality of information they are receiving. This is expressed as uncertainty or the desire for further validating or refuting data and further decision-making on how to factor that uncertainty into their assessment decision. A significant piece of validating data is hearing the “right” diagnosis, or the diagnosis in agreement with their own thinking from the student. Once a final assessment decision has been made, it must be communicated to the student along with guidance on how to address issues or improve, similar to a physician providing a patient with appropriate treatment after making a diagnosis. This process does vary slightly from hypothetico-deductive reasoning in that faculty enter encounters with patients with a number of different potential diagnoses in mind. When assessing a student it ultimately an assessment of whether the student did well and met specific criteria or did not.

Capitalizing on the similarities between the faculty reasoning process and the clinical reasoning process may provide an interesting opportunity to make novice-level clinical reasoning processes and good assessment practices more relevant and understandable to physician-assessors. A comprehensive assessment of clinical reasoning captures the students moving through different phases of the clinical process (i.e., information gathering, hypothesis generation, etc.). This process, however, can be difficult for more-experienced faculty to identify, both within learners and themselves. In the same way that expert physicians have difficulty making their reasoning process explicit to novice learners (Norman et al., 2007), I believe that physicians may have difficulty assessing the reasoning processes used by learners. Making this process explicit may help faculty understand the process they should use when completing assessments of clinical reasoning. In addition to teaching this process directly, pairing junior assessors with those with more experience can model good assessment practice.

Asking the senior faculty to vocalize their usually tacit thoughts around why they assessed the student the way they did offers a form of cognitive apprenticeship (Brown et al., 1989). This can be further enhanced through situated cognition if done within the setting in which future assessments will take place (Brown et al., 1989).

We also must ensure that faculty are not making decisions about student's ability prior to receiving all the information they need to make a sound assessment. This also parallels with clinical reasoning in that physicians should not decide on a diagnosis before collecting all the information, or be unwilling to change their diagnosis when new information becomes available, a phenomenon called premature closure. In assessment we see this with anchoring, where assessors initial judgments impact their ultimate decisions about the student (Epley & Gilovich, 2001). One way to prevent this is by providing clear scales with clearly defined behaviorally-focused anchors for faculty use to orient their initial assessment decisions (Crossley et al., 2011). Scales for assessing clinical skills have been used widely in medical education, and more recently scales have been developed for diagnostic reasoning that can be built upon and tailored to specific encounters (Society to Improve Diagnosis in Medicine, 2021).

Feedback

The final step within the model is feedback. The feedback a faculty member provides is directly influenced by the judgment, or judgements, they have made about the student. Because we now know that there is significant variability in how assessors arrive at their judgments and the judgments themselves, we would expect a similar level of variability across the feedback provided by each assessor. In addition to variability in content, however, faculty also take different approaches in the delivery of their feedback to learners. These differences included differences in addressing the student directly (e.g., “You did a great job…”) or indirectly (e.g., “The student did a great job…”) and using more passive vs direct language when providing constructive feedback.

In a standardized exam, such as an OSCE like the one in the encounter, it is not uncommon for a student to receive feedback from just the standardized patient or a single faculty member. In which case variation across faculty members would not be impactful on any individual student, unless that faculty members assessment is far outside the norm of how others are assessing or outside the criteria of the student should be assessed. In thinking through how this could have implications on assessments conducted in the real clinical setting, however, a student may interact with a dozen different faculty and residents in a four-week rotation who may all provide them with feedback. If every piece of feedback the student received was phrased or oriented differently, processing all that feedback, particularly in a small window of time, may be more challenging. Creating more consistency around feedback can also help the learner identify steps for improvement. Faculty’s varying language choices, particularly the use of more passive verbs and/or a suggestive tone, may make it difficult for students to evaluate which aspects of the feedback are important and “must-dos” as opposed to simply suggestions. Clear specific language leaves much less room for student interpretation and is appreciated by learners (Bienstock et al., 2007).

One way to create this consistency is to provide faculty development on how to provide assessment to learners. While the focus of this study was on how faculty make assessment decisions based on direct observation that required feedback to learners, it is also important to note that feedback is not always incorporated into every assessment, or may not be required or monitored. I believe this can significantly impact the quality of the feedback provided.

Together these different themes depict a process of faculty reasoning that can help us better understand how faculty make assessment decisions of students clinical reasoning ability. In addition, understanding the implications of the variation that exists across assessors, and the influence of context-specific factors, can inform institutions administering assessments on potential ways to mitigate variability that can influence assessment reliability.

VI. Conclusion

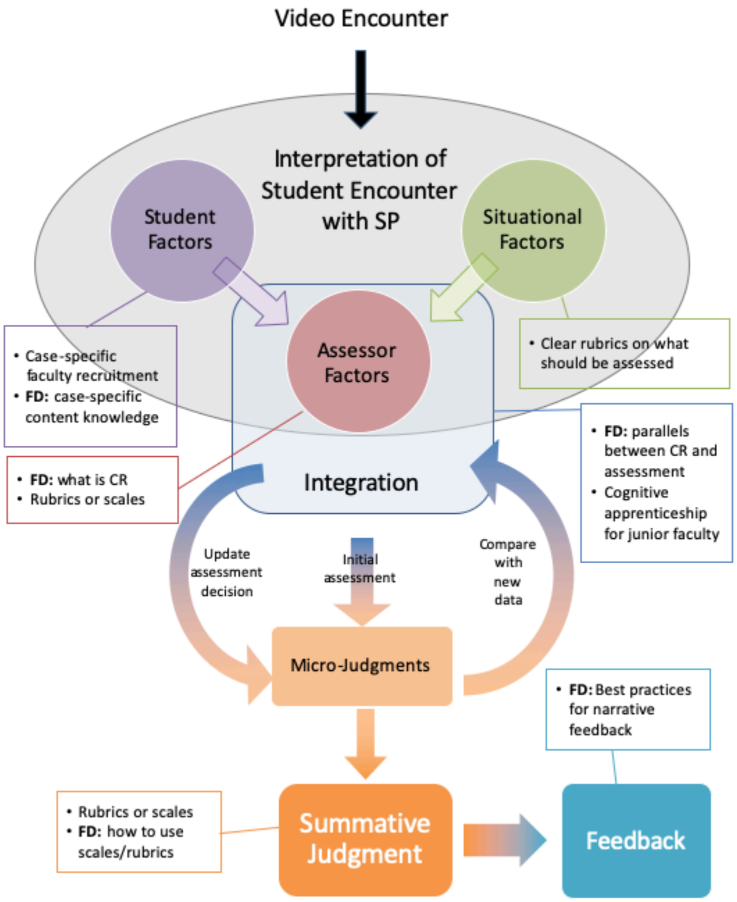
Through the use of think-aloud interviews and thematic analysis, I have identified a number of factors that influence the assessments faculty members make around clinical reasoning. Furthermore, I have identified one possible explanation of how the factors combine into a framework that outlines the reasoning process. Identifying these factors, and integrating them in a way that elucidates the internal, tacit process of faculty reasoning, provides opportunities to influence assessments and improve the reliability and validity of our assessments.

Opportunities for Influencing Assessments

When we step back and look at the faculty reasoning framework as a whole we can see opportunities to influence and mitigate different components of the reasoning process, with the goal of improving assessment validity. As noted in Figure 4, the most notable way in which we can improve assessments of clinical reasoning is through faculty development (FD), which is the process of providing training or education to faculty members participating in the education or assessment of students.

Figure 4

Opportunities to influence the faculty reasoning process


Video Encounter
Interpretation of Student Encounter with SP
Student Factors
Situational Factors
Assessor Factors
• Case-specific faculty recruitment
• FD: case-specific content knowledge
• Clear rubrics on what should be assessed
• FD: what is CR
• Rubrics or scales
• FD: parallels between CR and assessment
• Cognitive apprenticeship for junior faculty
Integration
Update assessment decision
Initial assessment
Compare with new data
Micro-Judgments
• FD: Best practices for narrative feedback
• Rubrics or scales
• FD: how to use scales/rubrics
Summative Judgment
Feedback

Faculty development is widely used in medical education, particularly around the teaching of clinical skills. However, historically, there has been far less faculty development occurring specifically around clinical reasoning and assessment (Holmboe et al., 2011). This could be reflective of the underrepresentation of explicit clinical reasoning curricula in medical schools in the United States (Rencic et al., 2017). Focused faculty development in the following areas could enhance the validity and reliability of clinical reasoning assessment:

- **Clinical Reasoning:** what clinical reasoning is; the different stages of clinical reasoning; how students should engage in the clinical reasoning process at different points in their training; and how to assess clinical reasoning in the context of the given assessment

- **Case-specific knowledge:** content-specific training about the subject matter of the case; expectations of what level of biomedical and clinical knowledge is expected of the learner; matching faculty specialty to case content
- **Assessment expectations:** rubrics or guidelines of what should be assessed; behavioral indicators to provide a frame of reference for student performance

Building upon existing literature

My framework of faculty reasoning also has similarities to past work around clinical reasoning and assessment. Gingerich et al. (2014) noted that “Research increasingly suggests that assessor expertise resembles diagnostic expertise in the clinical domain to a remarkable extent” (pg. 1061). Durning et al. (2013) identified a model of clinical reasoning using situated cognition that argues that clinical reasoning is influenced by three factors: physician factors, patient factors, and practice factors. The similarity between these factors and the ones I identified within the data strengthen the argument that the faculty reasoning process is similar to the process physicians use to diagnose patients, a process that is influenced by a number of factors.

Kogan et al. (2011) identified a model for the process assessors use in the direct observation of clinical skills with residents. Their model demonstrates similar factors as Durning et al. (2013) above: the patient, the trainee (in place of physician), and the clinical system and culture. In addition, similar to my model, they identify faculty assessors as significant factors, with similar components such as inference and frame of reference. One contrast to my model is the nesting of the factors within the culture and clinical system, which could be related to the situational factors I identified, although somewhat different. What I define as integration, Kogan et al. (2011) identify interpretation and synthesis.

The work of Kogan et al. (2011) provided valuable insight into how faculty members make decisions of residents, however, the data was collected in semi-structured interviews after the participants watched multiple encounters. The data for my study was conducted in real-time through the think-aloud process, and therefore perhaps better reflects the true and evolving thought process faculty use as they assess students. In addition, my work focused on the specific assessment of clinical reasoning, while past studies have focused on clinical skills, which represents a broader skill set. Because of these differences, I believe my study elucidated a deeper level of insight into how faculty navigate assessment decisions and grapple with the various pieces of information and information they are presented with in an encounter.

Future Directions

The data collected in my study is representative of a limited sample of faculty members at a single institution. I am encouraged, however, by the similarity in my findings with those who have conducted research around clinical reasoning and assessment at other institutions across the country. However, understand if the results of my study are truly generalizable requires significantly more research, which I hope to have the opportunity to do in the future. One direction I could see taking this work is to do a similar study with faculty members across more institutions. This would provide more generalizability to my findings as it could indicate if my framework holds true with faculty working outside of the UC Davis School of Medicine.

My study also asked faculty to respond to a single student. This was helpful as it held the stimulus constant and allowed me to identify similarities and differences that would only have come from the assessors themselves. The data from watching only a single student does, however, represents just that, how faculty would respond in a single assessment situation. Asking faculty members to watch and react to more than one student or more than one scenario. This could allow me to better understand how the faculty reasoning process varies when student factors or situational factors vary.

I would also like to use a similar methodology on materials gathered from the real-world clinical setting (i.e., video recordings of students interacting with real patients in the hospital or clinic). While the case I used is crafted from input of a number of different faculty members across several institutions, it is still a formulated case using a standardized-patient. Using a real patient encounter could change the dynamic between the assessor, student, and situational factors and therefore change the faculty reasoning process. I have hypothesized ways in which my findings could be applicable to assessment that occurs in the hospital or clinic, but the findings may not apply across contexts.

Lastly, the Zoom format of the encounter prevented the student from completing the encounter in the way that they normally would, and also prevented the participants from being able to interact with the student. Often faculty members will follow students' presentations with questions to dive deeper into their reasoning, which was not possible with this format. Finding a way to capture how those interactions affect assessment may be very interesting, and perhaps possible through direct observation of faculty-student interactions in the clinical setting.

Limitations

As noted above, one limitation of my study is that it includes a small cohort of faculty members from a single medical school. While a small sample size is acceptable for qualitative work, it does limit the generalizability of my findings. In addition, faculty members from other medical schools, where clinical reasoning is perhaps discussed more explicitly, may differ from UCDSOM faculty in how they make assessment decisions. As previously mentioned, applying the faculty reasoning framework developed in this study to TA transcripts collected at other institutions would be an exciting next step to determine generalizability. Although this is a clear limitation, I would also argue that the because faculty members in my sample did not all train (complete medical school and residency) at the same institution, it may be less of an impact on my findings. This may be more true of the more junior faculty who have been engrained in the UCDSOM culture for less time.

An additional limitation comes from the premise of competency-based medical education and reliable assessment, which is the need for repeated observations. Participants in my study were asked to make assessment decisions of students after only observing that student in a single encounter. While I think this holds relevance to some real-life situations in which faculty members must make assessment decisions after a single day, it is possible that faculty alter their decision-making process if they are relying on information from multiple encounters rather than just one.

Lastly, and perhaps most importantly, qualitative research is heavily influenced by the positionality of the researcher. My familiarity with some of the faculty participants may have influenced how I interpreted the data, however, this was avoided as much as possible by conducting the analysis with deidentified transcripts. Additionally, some for this study I was the only person who created, applied, and modified the codes and themes. While I validated some of my initial findings with a physician advisor to the project, my interpretations of the data and own inferences could have heavily influenced my findings.

VII. References

Ansari, A., Ali, S. K., & Donnon, T. (2013). The construct and criterion validity of the mini-CEX: a meta-analysis of the published research. *Academic Medicine*, *88*(3), 413–420.

Bandalos, D. L. (2018). *Measurement theory and applications for the social sciences*. New York: Guilford Press.

Bienstock, J. L., Katz, N. T., Cox, S. M., Hueppchen, N., Erickson, S., & Puscheck, E. E. (2007). To the point: medical education reviews-providing feedback. In *American Journal of Obstetrics and Gynecology* (Vol. 196, Issue 6, pp. 508–513). Mosby. https://doi.org/10.1016/j.ajog.2006.08.021

Boyd, P., Orr, S., & St, Y. (2009). *Grading Student Work: Using think aloud to investigate the assessment practices of university lecturers*. http://www.leeds.ac.uk/educol/documents/187564.pdf

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Brown, J. S., Collins, A., & Duguid, P. (1989). Situated cognition and the culture of learning. *Educational Researcher*, *18*(1), 32–42.

Coderre, S., Mandin, H., Harasym, P. H., & Fick, G. H. (2003). Diagnostic reasoning strategies and diagnostic success. *Medical Education*, *37*(8), 695–703.

Collins, A., & Gentner, D. (1987). How people construct mental models. *Cultural Models in Language and Thought*, *243*.

Crossley, J., Johnson, G., Booth, J., & Wade, W. (2011). Good questions, good answers: Construct alignment improves the performance of workplace-based assessment scales. *Medical Education*, *45*(6), 560–569. https://doi.org/10.1111/j.1365-2923.2010.03913.x

Daniel, M., Rencic, J., Durning, S., Holmboe, E., Santen, S., Lang, V., Ratcliffe, T., Gordon, D., Heist, B., Lubarsky, S., Estrada, C., Ballard, T., Artino, A., Sergio Da Silva, A., Cleary, T., Stojan, J., & Gruppen, L. (2019). Clinical Reasoning Assessment Methods: A Scoping Review and Practical Guidance. *Academic Medicine : Journal of the Association of American Medical Colleges*. https://doi.org/10.1097/ACM.0000000000002618

Deterding, N. M., & Waters, M. C. (2018). Flexible Coding of In-depth Interviews: A Twenty-first-century Approach. *Sociological Methods and Research*. https://doi.org/10.1177/0049124118799377

Donaldson, M. S., Corrigan, J. M., & Kohn, L. T. (2000). *To err is human: building a safer health system* (Vol. 6). National Academies Press.

Durning, S. J., Artino, A. R., Schuwirth, L., & van der Vleuten, C. (2013). Clarifying Assumptions to Enhance Our Understanding and Assessment of Clinical Reasoning. *Academic Medicine*, *88*(4), 442–448. https://doi.org/10.1097/ACM.0b013e3182851b5b

Dwyer, C. P., Hogan, M. J., & Stewart, I. (2014). An integrated critical thinking framework for the 21st century. *Thinking Skills and Creativity*, *12*, 43–52.

ECRI. (2020). *Diagnostic Errors, Maternal Health Top ECRI's 2020 Patient Safety Concerns*. ECRI Website. https://www.ecri.org/press/diagnostic-errors-maternal-health-top-ecri-2020-patient-safety-concerns

Elstein, A. S. (1994). What Goes Around Comes Around: Return of the Hypothetico-Deductive Strategy. *Teaching and Learning in Medicine*, *6*(2), 121–123. https://doi.org/10.1080/10401339409539658

Elstein, A. S., Kagan, N., Shulman, L. S., Jason, H., & Loupe, M. J. (1972). Methods and theory in the study of medical inquiry. *Academic Medicine*, *47*(2), 85–92.

Ennis, R. H. (1993). Critical thinking assessment. *Theory into Practice*, *32*(3), 179–186. http://www.tandfonline.com/doi/pdf/10.1080/00405849309543594

Epley, N., & Gilovich, T. (2001). Putting adjustment back in the anchoring and adjustment heuristic: Differential Processing of Self-Generated and Experimenter-Provided Anchors. *Psychological Science*, *12*(5), 391–396. https://doi.org/10.1111/1467-9280.00372

Ericsson, A., & Simon, H. A. (Herbert A. (1993). *Protocol analysis verbal reports as data*. MIT Press.

Ericsson, K. A. (2004). Deliberate Practice and the Acquisition and Maintenance of Expert Performance in Medicine and Related Domains. *Academic Medicine*, *79*(Supplement), S70–S81. https://doi.org/10.1097/00001888-200410001-00022

Ericsson, K. A., & Simon, H. A. (1998). How to study thinking in everyday life: Contrasting think-aloud protocols with descriptions and explanations of thinking. *Mind, Culture, and Activity*, *5*(3), 178–186. https://doi.org/10.1207/s15327884mca0503_3

Facione, P. (1990). *Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction (The Delphi Report).*

Fletcher, L., & Carruthers, P. (2012). Metacognition and reasoning. *Phil. Trans. R. Soc. B*, *367*(1594), 1366–1378.

Fonteyn, M. E., Kuipers, B., & Grobe, S. J. (1993). A Description of Think Aloud Method and Protocol Analysis. *Qualitative Health Research*, *3*(4), 430–441. https://doi.org/10.1177/104973239300300403

Funkesson, K. H., Anbacken, E.-M., & Ek, A.-C. (2007). Nurses' reasoning process during care planning taking pressure ulcer prevention as an example. A think-aloud study. *International Journal of Nursing Studies*, *44*, 1109–1119. https://doi.org/10.1016/j.ijnurstu.2006.04.016

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, *7*(2), 155–170.

Gingerich, A., Kogan, J., Yeates, P., Govaerts, M., & Holmboe, E. (2014). Seeing the 'black box' differently: assessor cognition from three research perspectives. *Medical Education*, *48*(11), 1055–1068. https://doi.org/10.1111/medu.12546

Gu, Y. (2000). To code or not to code: Dilemmas in analysing think-aloud protocols in learning strategies research. *IEEE International Symposium on Information Theory - Proceedings*, *43*, 236. https://doi.org/10.1016/j.system.2013.12.011

Hattie, J., & Timperley, H. (2007). The power of feedback. In *Review of Educational Research* (Vol. 77, Issue 1, pp. 81–112). Sage PublicationsSage CA: Thousand Oaks, CA. https://doi.org/10.3102/003465430298487

Higgs, J., Jones, M. A., Loftus, S., & Christensen, N. (2018). *Clinical reasoning in the health professions*. Elsevier Health Sciences.

Holmboe, E. S., Ward, D. S., Reznick, R. K., Katsufrakis, P. J., Leslie, K. M., Patel, V. L., Ray, D. D., & Nelson, E. A. (2011). Faculty Development in Assessment: The Missing Link in Competency-Based Medical Education. *Academic Medicine*, *86*(4), 460–467. https://doi.org/10.1097/ACM.0b013e31820cb2a7

Jackson, J. L., Kay, C., Jackson, W. C., & Frank, M. (2015). The Quality of Written Feedback by Attendings of Internal Medicine Residents. *Journal of General Internal Medicine*, *30*(7), 973–978. https://doi.org/10.1007/s11606-015-3237-2

Johnson-Laird, P. N. (1999). DEDUCTIVE REASONING. *Annual Review of Psychology*, *50*(1), 109–135. https://doi.org/10.1146/annurev.psych.50.1.109

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

Kogan, J. R., Conforti, L., Bernabeo, E., Iobst, W., & Holmboe, E. (2011). Opening the black box of clinical skills assessment via observation: A conceptual model. *Medical Education*, *45*(10), 1048–1060. https://doi.org/10.1111/j.1365-2923.2011.04025.x

Leighton, J. P. (2017). *Using think-aloud interviews and cognitive labs in educational research*. Oxford University Press.

Lundgrén-Laine, H., & Salanterä, S. (2010). Think-Aloud Technique and Protocol Analysis in Clinical Decision-Making Research. *Qualitative Health Research*, *20*(4), 565–575. https://doi.org/10.1177/1049732309354278

Merriam, S. B., & Tisdell, E. J. (2015). *Qualitative research: A guide to design and implementation*. John Wiley & Sons.

Norcini, J., Anderson, B., Bollela, V., Burch, V., Costa, M. J., Duvivier, R., Galbraith, R., Hays, R., Kent, A., & Perrott, V. (2011). Criteria for good assessment: consensus statement and recommendations from the Ottawa 2010 Conference. *Medical Teacher*, *33*(3), 206–214.

Norcini, J., Anderson, M. B., Bollela, V., Burch, V., Costa, M., Duvivier, R., Hays, R., Palacios Mackay, M., Roberts, T., & Swanson, D. (2018). 2018 Consensus framework for good assessment. *Medical Teacher*, *40*(11), 1102–1109. https://doi.org/10.1080/0142159X.2018.1500016

Norcini, J., Blank, L., Duffy, F., & Fortna, G. (2003). The mini-CEX: a method for assessing clinical skills. *Annals of Internal Medicine*, *138*(6), 476–481.

Norcini, J., & Burch, V. (2007). Workplace-based assessment as an educational tool: AMEE Guide No. 31. *Medical Teacher*, *29*(9–10), 855–871.

Norman, G. (2005). Research in clinical reasoning: past history and current trends. *Medical Education*, *39*(4), 418–427. https://doi.org/10.1111/j.1365-2929.2005.02127.x

Norman, G., Young, M., & Brooks, L. (2007). Non-analytical models of clinical reasoning: The role of experience. *Medical Education*, *41*(12), 1140–1145.

Oxford English Dictionary. (2020). *Oxford English Dictionary*. www.oed.com

Pangaro, L. (1999). A new vocabulary and other innovations for improving descriptive in-training evaluations. *Acad Med*, *74*(11), 1203–1207.

Pangaro, L. (2012). A Primer of Evaluation Terminology: Definition and Important Distinctions in Evaluation. In *ACE Guidebook for Clerkship Directors*.

Pangaro, L., & ten Cate, O. (2013). Frameworks for learner assessment in medicine: AMEE Guide No. 78. *Medical Teacher*, *35*(6), e1197–e1210.

Pottier, P., Hardouin, J., Hodges, B. D., Pistorius, M., Connault, J., Durant, C., Clairand, R., Sebille, V., Barrier, J., & Planchon, B. (2010). Exploring how students think: a new method combining think-aloud and concept mapping protocols. *Medical Education*, *44*(9), 926–935.

Ramani, S., & Krackov, S. K. (2012). *Twelve tips for giving feedback effectively in the clinical environment*. *34*, 787–791. https://doi.org/10.3109/0142159X.2012.684916

Rencic, J., Trowbridge, R. L., Fagan, M., Szauter, K., & Durning, S. (2017). Clinical reasoning education at US medical schools: results from a national survey of internal medicine clerkship directors. *Journal of General Internal Medicine*, *32*(11), 1242–1246.

Rojek, A. E., Khanna, R., Yim, J. W. L., Gardner, R., Lisker, S., Hauer, K. E., Lucey, C., & Sarkar, U. (2019). Differences in Narrative Language in Evaluations of Medical Students by Gender and Under-represented Minority Status. *Journal of General Internal Medicine*, *34*(5), 684–691. https://doi.org/10.1007/s11606-019-04889-9

Ryan, K. (2002). Assessment validation in the context of high-stakes assessment. *Educational Measurement: Issues and Practice*, *21*(1), 7–15. https://doi.org/10.1111/j.1745-3992.2002.tb00080.x

Saldaña, J. (2016). *The coding manual for qualitative researchers*. sage.

Schuwirth, L. (2002). Can clinical reasoning be taught or can it only be learned? In *Medical Education* (Vol. 36, Issue 8, pp. 695–696). John Wiley & Sons, Ltd (10.1111). https://doi.org/10.1046/j.1365-2923.2002.01274.x

Schuwirth, L., & Van der Vleuten, C. (2011). Programmatic assessment: from assessment of learning to assessment for learning. *Medical Teacher*, *33*(6), 478–485. https://doi.org/10.3109/0142159X.2011.565828

Society to Improve Diagnosis in Medicine. (2021). *Assessment of Reasoning Tool*. Society to Improve Diagnosis in Medicine. https://www.improvediagnosis.org/art/

Surry, L. T., Torre, D., & Durning, S. J. (2017). Exploring examinee behaviours as validity evidence for multiple-choice question examinations. *Medical Education*, *51*(10), 1075–1085.

Thampy, H., Willert, E., & Ramani, S. (2019). Assessing Clinical Reasoning: Targeting the Higher Levels of the Pyramid. *J Gen Intern Med*, *34*(8), 1631–1637. https://doi.org/10.1007/s11606-019-04953-4

Vu, N. V, Barrows, H. S., Marcy, M. L., Verhulst, S. J., Colliver, J. A., & Travis, T. (1992). Six years of comprehensive, clinical, performance-based assessment using standardized patients at the Southern Illinois University School of Medicine. *Academic Medicine*, *67*(1), 42–50. http://journals.lww.com/academicmedicine/Fulltext/1992/01000/Six_years_of_comprehensive,_clinical,.9.aspx

Wilby, K. J., Dolmans, D. H. J. M., Austin, Z., & Govaerts, M. J. B. (2019). Assessors' interpretations of narrative data on communication skills in a summative OSCE. *Medical Education*, *53*(10), 1003–1012. https://doi.org/10.1111/medu.13924

Young, J. Q., Van Merrienboer, J., Durning, S., & Ten Cate, O. (2014). Cognitive Load Theory: Implications for medical education: AMEE Guide No. 86. *Medical Teacher*, *36*(5), 371–384. https://doi.org/10.3109/0142159X.2014.889290

Young, M., Thomas, A., Lubarsky, S., Ballard, T., Gordon, D., Gruppen, L. D., Holmboe, E., Ratcliffe, T., Rencic, J., Schuwirth, L., & Durning, S. J. (2018). Drawing boundaries: The difficulty in defining clinical reasoning. *Academic Medicine*, *93*(7), 990–995. https://doi.org/10.1097/ACM.0000000000002142

VIII. Appendices

Appendix A

Recruitment Email

Good morning,

As you may already know, I'm currently completing my PhD in Education at UC Davis with a focus on medical education assessment and measurement. I'm excited to finally be in the dissertation phase of my study and am hoping for your assistance in recruiting my study participants, SOM clinical faculty with some experience working with medical students.

Please see the attached Letter of Information (also pasted below) for more information about the study. This study design was deemed "exempt" by the UC Davis Institutional Review Board.

Title of study: Understanding Faculty Assessment Decisions of Medical Student Clinical Reasoning Ability

Investigator: Marjorie Westervelt, MPH

Introduction and Purpose
You are being invited to join a research study. If you agree to participate in this research, you will be asked to participate in a think-aloud interview, which involves verbalizing your thoughts as you watch a video of a student in an objective structured clinical examination (OSCE), or in a focus group. Your participation in this research should take about 1 hour, which can be planned around your schedule.

When you participate in this research you will be video-recorded. The recording will be transcribed, but your name will not be included on the transcription.

Participation in research is completely voluntary. You are free to decline to take part in the project. You can decline to answer any questions and you can stop taking part in the project at any time. Whether or not you choose to participate, or answer any question, or stop participating in the project, there will be no penalty to you or loss of benefits to which you are otherwise entitled.

Questions
If you have any questions about this research, please feel free to contact the investigator at 415-602-0242 or mjwestervelt@ucdavis.edu.

If you or one of your colleagues would be willing to participate, please email me at mjwestervelt@ucdavis.edu to schedule a 1-hour time slot to complete your think-aloud interview.

If you have any questions please do not hesitate to reach out. Thank you for your consideration, I greatly appreciate it.

Appendix B

Door Note and Physical Exam Findings from the Encounter

Door note for student as entering the encounter:

Hannah Wesley is a 22-year-old female who presents with "stomach pains" in the Emergency Department.

Vital signs:
Temperature: 101.8 F
Pulse: 104/min
Respiration: 22/min
Blood pressure: 120/72 mm/Hg

You are to:
Take a relevant history.
Perform an appropriate physical exam.
Tell the patient what you think is going on with her and what you want to do next.
You will have 20 minutes to perform these tasks.

Pelvic Exam Findings:

The external genitalia are normal. There is cervical motion tenderness. The cervical OS is friable, with a yellowish discharge. Extreme tenderness on palpation of the right adnexa. Generalized tenderness on examination of the left adnexa. The uterus is not enlarged.

Appendix C

Questions after first pass of coding

During my first pass of coding I created 39 codes. I think some of these codes need to be refined and have the potential to be merged. I used primarily initial coding (Saldana, pg. 115) and concept codes (Saldana, pg. 119)

After reading, coding, and writing analytical memos for all of the transcripts my first reaction is the variability across the different interviews in what the faculty participants focused on and how they approached "assessment." Several questions related to my research questions have emerged that I'd like to explore and do some literature review on as i move forward in my analysis:

1) When and how do faculty make assumptions about student knowledge/reasoning versus making statements about observable knowledge or behaviors?
    a. Many of the faculty used the words “it seems” or “I think” or “I’m assuming” when trying to comment on how the student was thinking or why they were asking certain questions or making certain statements to the SP.
    b. There are other moments were faculty said “he’s asking X because of Y” and I’m curious to understand if that’s a safe statement to make (i.e. basic clinical knowledge or reasoning) or if those statements themselves are also assumptions.
    c. I’m interested to better understand how much these assumptions factor into the micro-judgments they make of students over the course of the encounter and how that leads to their overall assessment.
    d. I think these judgments could be interesting to compare with the work of Geoff Norman and his gestalt assessment work.
2) How does faculty forecasting future student behaviors affect their assessment decisions?
    a. Related to above, many faculty used terms of “I’m hoping they will…” and “Hopefully what’s happening in their head…” with a more optimistic tone that the student will get there or they are giving the student the benefit of the doubt.
    b. The code I used “Hope/Anticipate/Assume” needs some refinement but I think it captures a different vibe than the discussion above. Teasing out the assumptions from the optimistic hopes I think will be important as I look through and refine my first pass codes.
3) What comparisons are faculty making of the student’s performance: to themselves (what/how they would do it), to other learner levels (trying to identify the appropriate level), external checklists or practices (case checklists/scripts)?
    a. Faculty comparison of student to self: Many faculty used language of “what I would have done” or “what I always tell students to do.” This language was usually followed by a negative or neutral statement about something the student had just said or asked. It’s unclear to me as to whether or not these are best practices within medicine or personal preference from an experienced physician and if it’s “fair” to apply that standard to the student.
    b. Comparison to other learners: Many faculty compared the student’s performance to other students they had evaluated in the past in an effort to classify their level. I think this is natural in assessment, and because the student level wasn’t disclosed to participants until the end of the interview (if they asked) I think many faculty

members were trying to calibrate their judgments and feedback. I wonder if not providing level was a flaw, although I think I got some interesting data on how faculty grappled with making that judgment, and in real clinical practice where learners are all mixed together, it's not always apparent what level the student is and how to assess them.

   c. Comparison to external checklists: One faculty in particular wanted to compare the student to an external rubric or rubric in their mind. Other faculty made more general statements about "how students are taught" or "what we tell students to do" and comparing student performance to those checklists. I think this is interesting because many of the behaviors they referenced did not really align with any true reasoning process (i.e., washing hands) however they immediately honed in on these things. I also wonder about faculty (particularly newer faculty) relying on external rubrics, checklists, or information to assess clinical reasoning rather than having a true innate understanding. But perhaps that's an unrealistic expectation.

4) How much do faculty weigh their assessments on students getting the "right" answer vs the process the student used to arrive at their conclusion?
   a. Several faculty used language as "well he got there" or "in the end, he found the right differential" while others were more cautious, using language more like "if the signs were more subtle, would he have come to the right conclusion." This highlights to me the argument of is clinical reasoning only about coming to the right answer, or is the process truly what's important? And how do faculty measure or weigh that when assessing students.
   b. My impression is that faculty who were able to make comments about the reasoning and the differential along the way, are more able to see the process in action and that weighs into their overall assessment of the encounter. However, faculty who do not focus so much on this are really looking for the right answer. This could have interesting implications for faculty development.
5) What micro-judgments are faculty making throughout the encounter? How and when do they change? When do they remain consistent? How do they influence the summative judgment?
   a. Similar to bullet 4b. I noticed some students making micro or mini judgments about students reasoning throughout the encounter. Some would adjust their judgments accordingly as new information as presented, as if moving their rating up and down a sliding scale with each new piece of information. Some also used language that suggested they were making a judgment and had specific pieces of information they would be looking for in the future to verify or alter that judgment (which in itself is very similar to the clinical reasoning process!).
6) What do faculty miss that impacts their decisions? What is the impact of cognitive load?
   a. Multiple faculty asked me if students had asked a question or asked to go back to re-listen because they weren't sure if something was asked or missed it. They would also make comments of things the student missed, despite the student asking and receiving an answer from the SP on that information.
   b. Because faculty can't go back and re-watch an encounter in real-life and there isn't always the opportunity for them to verify with the patient what was said or done, this makes me think about how much faculty can really take-in and retain,

particularly when they are trying to keep track of what the patient is saying, what the student is doing, and make an assessment decision. Some faculty opted to take notes to help keep track, but even those who did take notes still missed pieces of information.

c. I believe there's been some literature on the cognitive load of SPs evaluating students on checklists, but I'd have to go back and check. And there has been plenty of research done on cognitive load in medicine in general. I'm not sure anything has been written about cognitive load of faculty particularly as it relates to assessment.

7) Are faculty focusing in on language and communication style assessing a different construct?

a. There was a group of faculty who gave much more emphasis on the language the student used, the types of questions, they asked, etc. as opposed to deeper level reasoning. As the interviewer who is also aware of the roles these individuals play in medical education and their experience level, it's hard for me to separate out this knowledge and not link it with this group, who I think are, in general, less experienced and tend to work with more novice learners (MS1s and MS2s).

b. While I would argue that communication is linked to clinical reasoning as it can affect data collection which is the first step (Daniels), this small group did not explicitly make that connection as to why they were focusing on it, unlike others who did comment that communication impacts how much data the patient will give the student.

c. As I read through these interviews I wonder if, even though prompted to assess clinical reasoning, they may have been assessing something else? Or maybe their understanding of what clinical reasoning is makes them focus in on communication? I think I need to explore this further to better interpret their transcripts.

8) How is the use of formulas/algorithms/checklists viewed by faculty? When is "systematic" a good thing or bad thing?

a. There was an interesting polarity I noticed with some faculty loving the use of checklists and algorithms because they were "efficient" or "ensured all the right data was captured" vs others saying this was too formulaic, didn't allow the patient to tell the story, should a lack of mastery, etc.

b. I think that there are different parts of the encounter that make these checklists more or less appropriate, and so I'd like to understand if this feedback was consistently applied (i.e. during the subjective its bad, but objective its good?).

c. I also wonder if the use of the checklist approach is more appropriate for certain levels of learners over others, and because the level of this student was not disclosed, could that be why there is misalignment?

9) Use of academic clinical reasoning lingo and other frameworks (RIME)

a. Some faculty used specific clinical reasoning terminology that others did not. This demonstrated to me a deeper understanding of the construct of clinical reasoning and what they should be paying attention to. These terms include metacognition, pertinent +/- , system 2 thinking, etc.

b. This also included applying other assessment frameworks in the absence of one being assigned to them. I believe 3 of the participants used the RIME framework without being suggested to. This shows a familiarity with medical education and assessment that beginner or non-educator faculty may not have.

   c. I have not yet gone through and coded for the Daniels et al clinical reasoning process codes, but it makes me think of coding those as "in vivo codes" (Saldana, pg. 105) to capture when specific terminology is used. I'll also include terms that are specifically defined or identified in the clinical reasoning literature as mentioned above.

10) How does faculty specialty influence their assessment of students? Do their expectations differ?
   a. Several faculty commented on how their specialty/training may influence their assessment of the student. Because this case had a significant gynecological component, the Ob/Gyn faculty member felt very comfortable with the content, but also felt that their expectations might be higher and could that be unfair. One PM&R doc commented that he really doesn't see cases like this regularly and that made him a bit uncomfortable in assessing the student. Another psychiatrist commented that he couldn't think of things the student was missing, but was sure an IM or surgeon could.
   b. These comments make me wonder how much the faculty members comfort level influences their assessment decision-making. In real practice, usually faculty assess students within their specialty and on cases they see routinely. However, in standardized environments we often recruit from a broad faculty base and take whoever is available. Ensuring faculty comfort level with the content could be really important in establishing expectations for assessment and an important note for faculty development. I also wonder how this plays out on more general specialties such as IM, EM, or FM where anything could walk through the door and its unrealistic for all faculty to be experts in everything.

11) How do faculty use context-specific data to alter their assessment of the student?
   a. A couple faculty members made remarks about how important the context (emergency room) was in how the student should behave and how that impacted their decision making on the student's performance. One faculty member in particular really struggled with how to assess the student because she felt that this was really not being taken into account. The language she used was very strong. Two others used the context to justify/explain why some of the questioning they did or what they "left out" (as opposed to missed) was appropriate. Some did not comment or asked for a reminder at the end and said "that was important" but did not expand. This information was provided to all participants at the beginning of the encounter in the Door Note.

12) How does the summative feedback vary across faculty? How does it differ from what is noted in the think-aloud?
   a. In addition to variability in the transcripts and feedback, there was also significant variability across the feedback that was provided. Because of the standardized encounter and that this feedback wouldn't actually get shared with the student, I hesitate to make any strong conclusions about how this feedback compares to what is provided to actual students. However, I do think there are differences in the types of language used, tone, and what comes out in the feedback as opposed to the transcript that is interesting to explore.

Appendix D

Final list of codes with definitions

#	Code	Freq.	Description
1	Integration		Faculty commenting on how they are interpreting the students actions or trying to make a determination about performance
2	Agreement	15	Faculty stated agreement with the student
3	Comparison to the "right" diagnosis	4	Faculty are assessing if the student has achieved the correct diagnosis
4	Determining what counts	9	Faculty are grappling with what is fair to include in their assessment, and what counts as clinical reasoning
5	Evaluation of reasoning process	40	Faculty are evaluating the students thinking/reasoning process and how they arrived at the diagnosis - not just the diagnosis itself
6	Interesting	24	Faculty stated "that's interesting" or "this is interesting" when referring to the students behavior.
7	Lack of knowledge	5	Faculty are assessing the students knowledge, or lack of thereof, and its impact on their performance
8	Uncertain/needs more info	22	Faculty are uncertain in their decision-making, may comment on wanting more information
9	Validating Data	14	Received or never received data that validated what they were thinking or expecting to get.
10	Feedback		Within the summative feedback, faculty shares perspective from first person context, rather than directive.
11	Active Constructive	9	When providing constructive feedback, faculty member uses a direct, active tone
12	Considerations for framing feedback	7	Faculty reflect on how they would frame feedback for learners and they would do so for this specific case
13	Direct	8	Feedback was directed at student. Use of "you"
14	Indirect	7	Feedback not addressed directly towards student. Use name of student or "student"
15	Passive Constructive	12	When providing constructive feedback, faculty uses more passive language
16	Suggest/recommend/consider / encourage/please	10	Faculty use words like consider, I recommend, or I suggest rather than giving direct instruction
18	Frame of Reference		Frame of reference participant uses to assess student - Kogan (2011)
19	External Frameworks	10	Faculty use an external checklist (or want one) or assessment frameworks when determing how to assess the student
20	Faculty to self	36	Faculty express what they would do in the scenario or would have liked to have seen. Comparing the student to themselves.
21	Student to others	70	Faculty compare the student or try to classify them to a certain level (i.e., M1, M4) or experience (novice, beginner). Often involves comparison to other learners at those levels.
22	Inference		Faculty are making an inference about the students performance - Kogan (2011)
23	Assumption	24	Faculty make an assumption about student performance or verbally say "I'm assuming"
24	Hope/Anticipate	27	Faculty optimism that the student will get there or waiting for them to ask a question or do something. More optimistic and positive in connotation.
25	Seems	14	Faculty use the word "seems" to denote uncertainty of what the student is actually thinking but they are trying to guess.
26	Think	32	Faculty use the word "I think" to comment on an opinion or thought about the student.

27	Judgment		Faculty state a judgment of the student's performance.
28	Micro-judgment	17	Faculty makes a micro-judgment of the students performance before the interview is complete.
29	Negative	87	Negative assessment of student
30	Neutral	5	Neither positive or negative assessment
31	Positive	113	Faculty likes or has a positive assessment of the student.
32	Summative	14	Faculty makes a summative judgment of the students performance
33	Knowledge of CR		Knowledge of the clinical reasoning process, familiarity with CR terminology
34	CR Terminology	42	Use of more academic clinical reasoning language (hypothesis, pattern recognition, pertinent +/-, illness scripts, etc.)
35	Illness Script	9	Faculty mentions the use of illness scripts or illness representations
36	Premature Closure	12	Faculty mentions premature closure or early anchoring
37	Daniel Stages of CR		The stages of clinical reasoning outlined by Daniel et al. (2020)
38	Daniel - Diagnostic Justification	3	
39	Daniel - Differential Diagnosis	70	the process of differentiating between two or more conditions which share similar signs or symptoms.
40	Daniel - Hypothesis Generation	2	
41	Daniel - Information Gathering	24	
42	Daniel - Leading or Working Diagnosis	0	
43	Daniel - Management and Treatment	6	
44	Daniel - Problem Representation	1	
45	Opinion of CR	21	Commentary on clinical reasoning from the faculty member (i.e. what is CR, what counts as being important to CR).
46	Situational Factors		Category grouping for factors that are unique to the given situation/context/case
47	Cognitive Load	17	Faculty missed something the student or SP said
48	Context	8	The context of the specific case is mentioned as a factor in decision-making
49	Specialty Influence	6	Faculty comment on how their specialty is influencing their judgments of the student.
50	Study Limitation	5	Faculty comment on limitations of the study that may inhibit their ability to assess the student (Zoom, can’t see their face, etc.)
51	Student Communication		Faculty noted specific language used by the student
52	Fillers	7	Student uses filler language (i.e., "Ums" "kind of") or unnecessary commentary
53	Funny/awkward/odd	14	Student uses funny, awkward, or weird language
54	Jargon	5	Student uses medical jargon to explain to patient
55	Missed opportunities	12	Faculty comment on a missed opportunity the student had to explain something or comment to the patient
56	Open/close ended questions	36	Faculty reference the student using an open or closed ended question or a leading question
57	Positive statements	4	Faculty comment on the students positive use of language or good statements
58	Terminology	3	Student used terminology that the faculty didn’t think appropriate when referring to the patient

59	Student Organization		Faculty comments on how the student organized the interview, including moving through HPI vs ROS, overall organization and when they chose to go broad vs focus in
60	Encounter flow	58	Organization or lack thereof of questions and how he moves through the encounter.
61	Focus	15	Narrow, hone in, or focus. Relates to questioning closer to pertinent information and differential or narrowing the differential itself.
62	Missing Info	61	Faculty note student is missing information, has gaps or holes in their data or plan.
63	Off-template	2	Faculty describe the student veering away from a template in a good way
64	Template/Algorithm/ Systematic	20	Faculty reference the student using a checklist, algorithm, or formula
65	Student-SP Interaction		Faculty note the student's general interaction with the SP
66	Body language	1	Comment on the student's body language
67	Patient Confidence	10	Patient's confidence in the student and their ability.
68	Patient Open-ness	3	Comment on the patient's openness to provide the student with information based on how the student is engaging with the patient
69	Pauses	7	Faculty commented on the student pausing.
70	Positive rapport	9	Faculty suggest the student is building a positive rapport, expressing empathy, that the patient appreciates something, etc.
71	Too much information	2	Faculty comment on the student overwhelming, providing too much information, or confusing the patient
72	Uncomfortable/nervous/ awkward	15	Comments on the student's discomfort, nervousness or awkward overall demeanor - not the same as use of specific awkward phrases or language