5-1-2019

Beyond Right or Wrong: The Influences of Thinking Disposition And Item Difficulty on Student Behavior During High-Stakes Testing

Kristina Lindquist
kristina.lindquist@tun.touro.edu

BEYOND RIGHT OR WRONG: THE INFLUENCES OF THINKING DISPOSITION AND ITEM DIFFICULTY ON STUDENT BEHAVIOR DURING HIGH-STAKES TESTING.

By
Kristina Lindquist

Bachelor of Science - Exercise Science
University of Massachusetts, Amherst
1999

Master of Science - Kinesiology
University of Nevada, Las Vegas
2006

A dissertation submitted in partial fulfillment
of the requirements for the

Doctor of Philosophy - Educational Psychology

Department of Educational Psychology and Higher Education
College of Education
The Graduate College

University of Nevada, Las Vegas
May, 2019

UNLV | GRADUATE COLLEGE

Dissertation Approval

The Graduate College
The University of Nevada, Las Vegas

April 12, 2019

This dissertation prepared by

Kristina Lindquist

entitled

Beyond Right or Wrong: The Influences of Thinking Disposition and Item Difficulty on Student Behavior During High-Stakes Testing.

is approved in partial fulfillment of the requirements for the degree of

Doctor of Philosophy - Educational Psychology
Department of Educational Psychology and Higher Education

Alice Corkill, Ph.D.
Examination Committee Chair

Kathryn Hausbeck Korgan, Ph.D.
Graduate College Dean

Lisa Bendixen, Ph.D.
Examination Committee Member

Harsha Perera, Ph.D.
Examination Committee Member

Merrill Landers, Ph.D.
Graduate College Faculty Representative

Abstract

A hallmark of clinician decision making is the ability to know when to make quick decisions and when decision making should be slowed to account for complicating factors. Throughout the physician training process, multiple choice test items are used to assess student knowledge however, these items do not assess the process used by a student to arrive at the answer choice. If an important characteristic of decisions in clinical practice is timing, then decision timing could be an important consideration for medical school assessments. The purpose of this study, therefore, is to investigate factors that may affect the amount of time a student takes to consider a test item when first presented with the item. Based on decision making theory, factors are at the student-level (current knowledge, prior knowledge, and thinking disposition), item-level (item difficulty), and student-by-item-level (response) should have an effect on the amount of time a student views a question on the first encounter.

The present study employed student testing data generated by second-year medical students during a 154-question high-stakes exam. First-view time of each item was derived from computer-based snapshot files and used as the dependent variable. A multilevel mixed model was specified to describe the effect of response, item difficulty, and student thinking disposition on first-view time. As predicted, students spent less time on first encounter for items which were ultimately answered correctly. However, for the easy items answered incorrectly, students spent the same amount of time as difficult items answered incorrectly. In addition, students who were more analytical in their thinking disposition displayed less difference in first-time view between correct and incorrect responses. These results are interpreted using the theoretical framework of the three-stage dual process model of decision making proposed by Pennycook, Fugelsang, and Koheler (2015).

Acknowledgements

While a dissertation is a demonstration of a student's ability to conduct individual research, it is not a document that is produced in solitude. Many individuals contributed to this work through direct and indirect contact. I would like to thank my doctoral committee chair, Dr. Alice Corkill for her encouragement, support, and countless edits during this process; my doctoral committee, Dr. Harsha Perera, Dr. Lisa Bendixen, and Dr. Merrill Landers for their support in the research process.

In addition, I would like to thank Dr. Anne Poliquin for her mentorship as a colleague and friend during the doctoral program. Dr. Terrence Miller and Mr. Redan Hablero greatly contributed to this work by developing the ExamSoft snapshot analysis tool used to convert the raw data from ExamSoft. Finally, I would like to thank the DO21 cohort and Touro University Nevada for their overwhelming participation in this study.

Table of Contents

List of Tables

List of Figures

Chapter 1: Overview

Contrary to popular belief, medical students are not a homogenous group of stellar students. Some medical students struggle immensely during their didactic training while others find this training relatively easy. Medical students, in this regard, are the same as any other student; their knowledge, skills, and abilities differ and, correspondingly, so does their performance during medical school.

Typically, students are assessed during the first two years of medical school with high-stakes, multiple-choice exams. These exams are high-stakes because each exam factors into their grade point average and class rank which, in turn, are used by residency programs to determine the desirability of a student for post-graduate medical training. Multiple-choice assessments are valuable because they allow faculty members to clearly assess whether a student understands the material presented. While multiple-choice exams provide an easy means for assessing binary student knowledge (i.e., the student either knows the answer or not), they do not provide insight into the level of a student's knowledge with regard to the concept tested by the item. For example, a student may understand the disease pathway associated with a specific set of symptoms but may miss a vital clue when thinking through the pathway. This clue may allow him/her to differentiate between two similar diseases; if both the correct disease and incorrect "close" disease are in the answer choices, the student who chooses the "close" disease receives an incorrect mark. Thus, the incorrect "close" response is equal to a completely incorrect response.

Students leave clues to their knowledge and understanding when they are taking exams and these clues can be discovered when they take computer-based exams. Information such as the amount of time a student takes when considering an item and the number of times a student

returns to a particular item can provide valuable information about the student's level of understanding. When these data are aggregated over an entire exam, they can provide information that faculty can use to determine the level of learning for individual students and for the class as a whole.

Research Purpose

The purpose of this study is to investigate factors that may affect the amount of time a student takes to consider a test item when first presented with the item. Based on decision making theory, factors are at the student-level (current knowledge, prior knowledge, and thinking disposition), item-level (item difficulty), and student-by-item-level (response) should have an effect on the amount of time a student views a question on the first encounter.

Previous research on student performance and test items has focused on the number and direction of change for student answers on multiple choice questions (Bauer, Kopp, & Fischer, 2007; Couchman, Miller, Zmuda, Feather, & Schwartzmeyer, 2016; Ferguson, Kreiter, Peterson, Rowat, & Elliott, 2002; Stylianou-Georgiou & Papanastasiou, 2017). In addition, multivariate models that incorporate response time into the ways that exam items function have been proposed (van der Linden, W. J., 2011; Wang & Hanson, 2005; Wise & Kong, 2005; Zhan, Jiao, & Liao, 2018), however these models focus on how the exam items perform rather than exploring what timing data can add to our understanding of student performance.

Clinical decision-making of practicing physicians and residents has been explored through behavioral patterns associated with answering multiple-choice questions (Eva & Regehr, 2007; McConnell, Regehr, Wood, & Eva, 2012) and recall of decision-making experiences (Moulton, Regehr, Lingard, Merritt, & MacRae, 2010). Data regarding medical student behavior during testing, beyond answer changing, is not apparent in the literature. The absence of

information related to medical student testing behaviors is because measurement of performance variables may impede students during high-stakes testing. However, students taking computer-based tests generate these data without realizing measurements are being conducted, therefore collecting and analyzing these data can provide insight into testing without interference.

Literature Review

The reasons why students generate incorrect answers during testing has long drawn the attention of researchers. Understanding how students typically generate incorrect answers on material they have (presumably) studied would benefit teachers and students at all levels. Research on this topic has approached the problem from two sides: 1) top down approaches that attempt to address the misconceptions that students have regarding the topic at hand and 2) bottom up approaches that explain the environmental variables that predispose a student to incorrect answers (Heckler, 2011).

Research examining misconceptions has revealed that students with misconceptions stored in memory may be subject to miscategorization of the material, particularly at the ontological level (Chi, 2005). In addition, if students have generated their own explanations of the phenomenon at hand, they may have missed key details in their explanation that would provide a deeper understanding of the material (Rozenblit & Keil, 2002). Finally, students may be motivated to retain a particular misconception – this may be due to the concepts they have learned through experience (Sinatra, Kienhues, & Hofer, 2014) or their epistemological beliefs (Hofer & Bendixen, 2012).

However, students often have the correct information stored in memory and, when questioned outside of a testing situation, are able to easily produce the correct answer (Heckler, 2011). Even experts are known to fall prey to this phenomenon. Shtulman and Valcarcel (2012) found that expert scientists provided incorrect answers about the Earth's orbit around the sun

when tested under timing and working memory load conditions. Therefore, incorrect answering may be due to the stress that students experience during testing situations rather than true memory misconceptions. The most common methods for investigating students reactions during testing are examinations of answer changing behavior (Benjamin, Cavell, & Shallenberger, 1984; Ferguson et al., 2002; Fischer, Herrmann, & Kopp, 2005; Geiger, 1996; Jeon, De Boeck, & van der Linden, 2017; Stylianou-Georgiou & Papanastasiou, 2017) and explorations of exam taker confidence (Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013; Jackson, Kleitman, Howie, & Stankov, 2016; Pennycook, Ross, Koehler, & Fugelsang, 2017; Szollosi, Bago, Szaszi, & Aczel, 2017). However, investigations of answer changing provide no insight into the reasons why students change their answers, and explorations of exam taker confidence are unreliable when taken at the end of an exam but intrusive when asked during exam performance.

Theoretical Framework

Pennycook, Fugelsang, and Koehler (2015) described a three-stage, dual-process model for decision making that could help describe students' behaviors during testing. In their model decisions can be made quickly and intuitively through heuristic processes (Type 1 thinking, T1, intuitive) or decisions can proceed slowly through reflective and analytic thought processes (Type 2, T2, analytic or reflective). This model builds on a tradition of dual-process thinking in decision making that was first described by Tversky and Kahneman (1974).

Tversky and Kahneman's (1974) original research explored how individuals reached incorrect conclusions through the use of mental short cuts called heuristics. Use of heuristics for decision making has typically been assessed with two manipulations of intuitive and analytic thinking. The first is the use of extreme base-rates, where individuals need to choose between answering an item based on statistical likelihood or stereotypical descriptions provided in the

text (Kahneman, 2011). The second is the Cognitive Reflection Test (CRT), a three- to eight-item assessment that employs questions that typically elicit an incorrect "intuitive" answer but can be correctly answered if the intuitive answer is inhibited in favor of a more analytic approach to the question (Frederick, 2005b; Primi, Morsanyi, Chiesi, Donati, & Hamilton, 2016; Toplak, West, & Stanovich, 2013).

Intuitive thinking. The three-stage, dual-process model of decision making employed here was developed using both base-rates and the CRT to inform the model (Pennycook et al., 2015). In this model, decision-making begins with intuitive answers being generated quickly when an individual is presented with a decision-making task. Type 1 thought processes are generally based on experience, therefore the generation of intuitive answers in a decision-making task is heavily influenced by an individual's previous experience (Gigerenzer & Brighton, 2009; Klein, Calderwood, & Clinton-Cirocco, 1986). Multiple intuitive responses may be generated through T1 thought processes and these responses can vary in their appropriateness for the decision-at-hand (Sloman, 1996). The answer expressed in this case would be the answer that requires the least amount of cognitive work to generate (Toplak et al., 2013). This may explain why students are more likely to generate incorrect responses during testing even when they have the correct answer stored in memory: their older "intuitive" answers are less cognitively taxing because they come to mind easier than their new "learned" answers.

Cognitive dissonance. If an intuitive answer is generated and expressed, it is generally thought that the answer does not go through a process of "checking" by slower, analytical T2 thought processes (Pennycook et al., 2015). However, sometimes an individual is primed for answer checking. This could occur through specific directions to "carefully consider answers before choosing" (Eva & Regehr, 2007) or it could be a personal disposition towards careful

consideration of answers (Jackson et al., 2016). When queried on which specific factors caused an individual to check an answer, research participants often report that there is no clear signal that causes reconsideration and that it is just a "gut feeling" that something is not right (Eva, Link, Lutfey, & McKinlay, 2010; Moulton et al., 2010).

While there may be no identifiable metacognitive signal to change from T1 to T2 processing, the conflict between multiple T1 responses has been posited as the reason for the switch to T2 processing (Pennycook et al., 2015; Sloman, 1996), and this conflict has been labeled "cognitive dissonance." At this point, individuals are able to consciously consider the intuitively generated responses and they can override the response through cognitive decoupling or they may support the response through rationalization (Pennycook et al., 2015).

Cognitive decoupling and rationalization. Type 2 responses require additional processing time. Cognitive decoupling is the process that an individual goes through when considering why an intuitive response may not fulfill the parameters of the decision-making task (Pennycook et al., 2015). In student testing data, this may be expressed by the student taking additional time when viewing an item, revisits to an item, or an answer change from incorrect to correct, particularly when the answer change occurs in conjunction with additional time or revisits to an item.

Rationalization is the process of supporting an intuitive response by reasoning through why the intuitive response is correct (Pennycook et al., 2015). In student testing data, this may appear as an item with an incorrect response that takes the student a particularly long time to answer, or an item that is revisited multiple times without an answer change.

Both correct and incorrect answers can be reached through T1 and T2 processes. The three-stage, dual-process model for decision-making offers four possible process/answer

combinations: Intuitive/Incorrect, Analytical/Incorrect, Analytical/Correct, and Intuitive/Correct (Pennycook et al., 2015). The determining factor between the generation of correct and incorrect answers is the individual's knowledge base, whereas the determining factor between the speed of answer generation could be the individual's thinking disposition (Jackson et al., 2016; Pennycook et al., 2015).

Research Questions

Students' behavior on high-stakes testing should follow the predictions of the three-stage, dual-process model of decision-making. Based on a student's knowledge of the tested material, answers to multiple choice questions could be generated through T1 or T2 processes. Incorrect answers generated through T1 processes indicate that the student has an intuitive misconception regarding the material tested (Intuitive/Incorrect). Incorrect answers generated through T2 process indicate rationalization of an incorrect T1 response (Analytic/Incorrect). Correct answers generated through T2 processes indicate cognitive decoupling, therefore the student has the correct concept in memory (Analytic/Correct). Correct answers generated through T1 processes indicate that the student has learned a concept to the level of intuitive responding (Intuitive/Correct).

A confounding factor in this analysis could be students' thinking disposition unrelated to knowledge, such that students who could generate Analytic/Correct answers may not do so because of a preference for intuitive answering (Pennycook et al., 2017). Conversely, some students who are capable of generating Intuitive/Correct answers may generate Analytic/Correct answers because they lack confidence or have a high confidence threshold for answer selection (Jackson et al., 2016).

The research question posed here was: Is there a relationship between the amount of time a student spends on the first-view of an item and the response ultimately generated by the student? If there is a relationship between response and first-view, do factors such as question difficulty, student knowledge, and/or thinking disposition affect this relationship? First-view time for a question is an important measure of a student's immediate knowledge of the question and is a reflection of the student's preferences with regard to answering and confidence. If a student generates a quick and correct answer, therefore, it can be inferred that the student is relatively knowledgeable about the concept addressed by the item and is confident in that knowledge. While additional visits and total question view time could also factor into considerations of knowledge and confidence, first-time view was chosen for this study because it was present for each item (unlike additional visits), and less likely to be confounded by

Research Design

Test taking behaviors of 108 second-year medical students were measured during a 154 item high-stakes, multiple-choice exam. Computer-based testing software (Exam Soft Inc., 2017) was used to deliver an examination. A by-product of this was the generation of a log file of exam snapshots. A snapshot was generated when a student began an item, when the student selected an answer, when the student left an item, if a student returned to an item, and if a student selected a different answer (on either the first or subsequent views). Through a data reduction procedure developed by Miller and Lindquist (2016), the snapshots yielded information regarding the amount of time a student had the item on screen during the initial encounter with the item (first-view time) and whether the item was answered correctly or incorrectly (response).

In conjunction with the student's test taking behavior, students' thinking disposition was assessed using the three-item Cognitive Reflection Test (CRT; Frederick, 2005) after completion

of the exam. The CRT used for assessing thinking disposition consisted of the original three-items proposed by Frederick (2005), however the four-option multiple-choice format was used to more closely mirror decision making during testing (Sirota & Juanchich, 2018). Student's thinking disposition (analytic/intuitive) was scored by the number of correct answers generated (Frederick, 2005)

A multilevel mixed model was constructed to determine the relationship between the amount of time a student spent on an item the first-time it was encountered, the response generated, the difficulty of an item, and person-level factors. The person-level factors in the model were score on the current exam (an indicator of current knowledge base), first-year grade point average (an indicator of previous knowledge base), and score on the cognitive reflection test (an indicator of thinking disposition).

Multilevel mixed models are built by answering four consecutive questions. For this research, the questions were:

1. Does first-time view vary across students, items, or students by items?
2. Does a student's response to an item (correct or incorrect) explain differences in first-time view? Does item difficulty (item-level factor) explain differences in first-time view? Do thinking disposition, current knowledge, or prior knowledge (person-level factors) explain differences in first-time view?
3. Does the item-level factor of difficulty explain variation in the relationship between response and first-time view? Do the person-level factors of current knowledge, prior knowledge, and/or thinking disposition explain the relationship between difficulty or response and first-time view?

4. Are there combinations of factors across levels that moderate the relationship between response and first-time view?

Limitations/Gaps

This study was limited in scope, participants in the study were from a single medical school cohort at one institution. While the results of this study may provide a window into student performance, these data are only an initial investigation of how students interact with exam items. In addition, there may be individual factors that affect student performance on exams beyond their knowledge base and thinking disposition.

Significance of the Study

Investigations of dual process theory have examined intuitive and analytical responses to syllogistic reasoning problems (Handley & Trippas, 2015), base rate problems (Pennycook, Trippas, Handley, & Thompson, 2015), CRT items (Travers, Roliston, & Feeney, 2016), and conditional reasoning tasks (Thompson, Prowse, Turner, & Pennycook, 2011). However, few, if any, studies have examined how performance on theoretical measures of dual process thinking are connected to decision making outside of experimental settings. This study was a first attempt to bridge the theoretical and applied worlds; to investigate thinking and decision making of students on high-stakes tests within the framework of the three-stage dual-process model.

This investigation advances the theoretical base by directly testing the hypotheses suggested by the three-stage dual process model proposed by Pennycook and colleagues (2015). In addition, the study provides a base for further investigations of student decision making by describing the relationship between an exam taking behavior, the amount of time spent considering an item on the first encounter, and person-level factors. This investigation could

prove fruitful to future theoretical investigations by proposing new variables that could further explain observations related to dual process thinking.

Clinical Relevance

The results of this study can also be viewed by their clinical significance – both to the educational/teaching profession and the medical profession. Teaching and education has long assessed students' knowledge along one domain, the ability to provide the correct answer, especially when responding to multiple choice exam items. In the medical field, didactic education struggles to inform clinical practice; often students who perform well in the classroom have trouble with clinical skills and vice versa. This study may begin to inform both the general education community and the medical education community of new ways to assess students' knowledge of taught material.

While students' tendencies to change answers on exams has been investigated in the past (Bauer et al., 2007; Benjamin et al., 1984; Couchman et al., 2016; Ferguson et al., 2002; Fischer et al., 2005; Geiger, 1996; Stylianou-Georgiou & Papanastasiou, 2017), the inclusion of question timing and analytic tendency represents a new method of examining student performance. The results of this study could be further extended to create a rubric for multiple choice exam performance where students may have developed expert level knowledge of a concept and therefore answer questions quickly and correctly; intermediate level knowledge where the student answers questions slowly, sometimes providing correct answers, but sometimes unable to override previously held conceptions of that knowledge; or novice level knowledge where a student is unlikely to detect their mistaken intuitive answers. This could provide teachers with a powerful diagnostic tool that enables more precise teaching; if the majority of a class slowly answers an item incorrectly, the teacher can surmise that the students detected a conflict between

their previously held beliefs and the learned information, perhaps these students would benefit from a study technique such as elaborative interrogation or self-explanation (Dunlosky et al., 2013). Conversely, if the whole class answers an item quickly with the intuitive distractor, perhaps the students need additional instruction to counter misconceptions in their knowledge base (Chi, 2005; Sinatra et al., 2014).

For medical educators, a large body of literature is developing regarding the clinical importance of "knowing when to slow down" (Chang, Kang, Ham, & Lee, 2016; Eva & Regehr, 2007; Hess, Lipner, Thompson, Holmboe, & Graber, 2015; Hruska, Krigolson et al., 2016; Moulton et al., 2010). Expert physicians do not approach all decisions as equal. They are able to know when they need to slow down based on the information they have (or do not have); in essence, clinicians who "know when to slow down" are more in tune with the conflict monitor in Pennycook and colleague's (2015) three-stage dual process model. If students could learn this tuning prior to entering the clinic, they could have more productive clinical experiences because they might understand when it is most appropriate for them to use intuition and when it is more appropriate to use T2 processes to solve a diagnostic or treatment problem.

Summary

Is there a relationship between testing behavior and the production of an answer? If there is a relationship between testing behavior and answer production, is the relationship moderated by item difficulty, student knowledge, and/or thinking disposition? The literature reviewed in chapter two indicates that there is an interaction between knowledge and thinking disposition when students respond to items during exams, but this interaction is not well understood. From the decision-making literature, a theoretical model that incorporates some aspect of speed may prove useful for analyzing student decision making during test taking. From the test taking

literature, more advanced students have been identified as having more nuanced approaches to test taking. Therefore, students who have progressed to higher levels of education may display more variability in testing behaviors. Finally, from the literature regarding student misconceptions, we can relate the speed of a decision with response patterns - correct responses that are generated quickly are concepts that are well-learned, incorrect responses generated quickly are most likely long held misconceptions. Conversely, slowly generated responses that are correct are concepts that are still being learned whereas slow, incorrect responses may indicate that an individual is debating between a misconception and the learned concept.

Chapter 2: Literature Review

In the song "The Gambler" (Rogers, 1978), Kenny Rogers delivers sage advice for poker players and test takers: "You have to know when to hold 'em, know when to fold 'em, know when to walk away, know when to run." However, this is not typically the advice that students receive regarding testing, especially for high-stakes, multiple-choice tests. In these situations, students often receive advice that only focuses on the application of a single strategy to all questions. Students are often advised to either "answer quickly and don't change answers" or "slowly consider each question and the range of answers available". Unlike the advice of "The Gambler" students are not urged to consider the context of their decisions during testing, but to stick to a single rule that will, presumably, maximize their chances for answering correctly. Not only does this advice indicate confusion about how students should approach multiple-choice questions, it presents an incomplete picture of what test takers are doing when they make decisions during high-stakes tests.

For the purpose of this literature review there are two major considerations for the behaviors of students during test-taking. Examining the incorrect responses that students provide during testing enables an investigation of the types of misconceptions that students hold in memory. These misconceptions, arguably, generate the incorrect answers given to exam items. The other consideration is that environmental factors within the testing environment, factors such as timing, item difficulty, and stress, affect how students respond to exam items. These "top down" and "bottom up" perspectives on testing yield salient clues to understanding the test taking process and, when considered in light of current theories of decision-making, may provide additional insight into a student's knowledge beyond simple right and wrong answers.

Test Taking Behaviors in Undergraduate Populations

Undergraduate students, particularly those studying science, have provided much of the data related to student performance patterns on multiple choice questions. Patterns of incorrect answers have been interpreted two ways: top-down cognitive processes that are incorrectly specified in memory (Chi, 2005; Rozenblit & Keil, 2002) or bottom-up cognitive processes that influence, and potentially override, correct answering (Heckler, 2011; Tversky & Kahneman, 1974).

Top-down cognition: Concepts, categories, and explanations. Ultimately, the generation of an incorrect response to an item on an exam displays a student's lack of knowledge associated with the intended knowledge concept. Misconceptions in memory have been attributed to incorrect ontological categorization of concepts (Chi, 2005), a shallow understanding of the concept that leads to confident, but incorrect, answers (Rozenblit & Keil, 2002), or motivated reasoning towards concepts already stored in memory (Sinatra et al., 2014).

Concepts. When "new" information is presented, it is likely that students already have some concept related to that information stored in memory. For example, children often have a concept of planetary movement prior to learning about the solar system in school. This concept or belief is untutored and typically based on observation. Hence, most individuals begin with a concept that the sun moves around the Earth based on the observation that the sun appears to rise in the east, traverse overhead, and set in the west (Carey, 2000; Keil, 2011). Learning the currently accepted scientific explanation for concepts, therefore, requires the "replacement" of old information with new information. However, replacing information in memory is difficult, in part because the older, incorrect concept holds a stronger memory trace, but also because each time the new concept is recalled, the old concept is also activated and inhibited (Shtulman &

Valcarcel, 2012). If the learner is unable to inhibit the older, stronger concept due to the imposition of cognitive load, the incorrect concept will be used to generate an incorrect answer. Even expert physicists, with their advanced understanding of planetary motion, will endorse the concept that the sun revolves around the Earth if they are placed under extreme cognitive load conditions (Shtulman & Valcarcel, 2012).

Categorization. To further understand misconceptions, it is necessary to examine the structure of concepts that are naïve. Chi (2005) proposes that robust misconceptions cross ontological categories. Ontological categories are the basic categories of reality or existence in the world. Each of these categories has specific attributes that are plausible for that category, but not for other categories. Therefore, robust misconceptions occur when plausible (perhaps even intuitive) attributes from one ontological category are assigned to another ontological category (Chi, 2005). Beyond the confusion of category assignments, some processes are more intuitive for students to understand. Chi (2005) calls these intuitively understood processes "direct processes" because students more naturally understand that the components of the system create the mechanical patterns created by the system. The converse of direct process, "emergent process", is less intuitive for students to comprehend. Here, students typically infer direct action of the components of the system without viewing the collective action of the system (Chi, 2005)

Direct process misconception may result in an explanation that is correct on the surface but is not stable when considered at a deeper level (Chi, 2005). For example, students typically explain the process of diffusion by stating that molecules move from an area of higher concentration to an area of lower concentration until equilibrium is reached. On the surface this explanation is correct, however, when asked how this movement occurs, students will typically indicate that the molecules first move from high concentration to low concentration and once

equilibrium is reached, the molecules cease movement. A complete explanation of this emergent process would incorporate the collective action of the system such that molecules can move in all directions (including from high to high concentration or low to high concentration), and that once equilibrium is achieved, movement continues (Chi, 2005).

As Chi's (2005) discussion of misconceptions indicates the structure of a student's explanation for a concept can provide valuable insight into the depth of their understanding. Explanations are a way that we, as humans, attempt to systematize the world around us, whether they relate to everyday occurrences or to scientific phenomena (McCain, 2015). Successful explanations adequately incorporate the data at hand, cover a range of associated data, have internal consistency, are consistent with current theory, convey a sense that the explanation is not simply "made up" for the current data, and have the ability to predict other (associated) phenomena (McCain, 2015). Explorations of the process of generating explanations have uncovered differences in the ways that novices and experts explain scientific phenomena (Rozenblit & Keil, 2002; Sherin, Krakowski, & Lee, 2012).

Explanations. Novice explanations are characterized by the inclusion of extraneous surface details that the individual believes to be important (Rozenblit & Keil, 2002). In addition, novice explanations fail to encompass one or more of the components of "successful" explanations because the individual does not truly understand the underlying mechanisms of the phenomenon at hand (Rozenblit & Keil, 2002). Due to the inclusion of many surface details and the lack of consideration for underlying (or unseen) mechanisms, novices tend to generate fast explanations that incorporate a variety of surface characteristics; this leads the novice to rate their explanations with high confidence despite the fact that the explanation generally fails to explain the phenomenon at hand (Rozenblit & Keil, 2002). Rozenblit and Keil's (2002)

experiments to explore the "Illusion of Explanatory Depth" found that novices to the phenomenon (regardless of their education level) generated explanations of crossbows, helicopters, and toilets that highly depended on surface characteristics such as color, material, and decoration when attempting to explain how these items functioned. Experts only incorporated these details when they were vital to the function of the item (e.g., the helicopter is made out of aluminum because it is a lightweight metal; Rozenblit & Keil, 2002).

When novices are confronted with the contradictions that appear within their self-generated explanations, they tend to take one of three paths to reconciling the contradiction. Some may change their explanation, in some cases completely dropping the old explanation in favor of a new, more encompassing explanation (Sherin et al., 2012). Others may attempt to "fit" the explanation around the contradiction without changing the original explanation concept. Finally, a small minority will reject the contradiction in order to maintain the explanation "as is" (Sherin et al., 2012).

The reasons why a student may adopt, fit, or reject a new explanation may reside in the student's epistemic beliefs and emotions related to the new explanation (Sinatra et al., 2014). An individual's epistemic beliefs relate to the nature, source, and limits of knowledge (Hofer & Bendixen, 2012) and affect how an individual assimilates new information into their knowledge base (Franco et al., 2012). Students who believe knowledge is certain and simple may rarely doubt the answer they provide on a multiple-choice exam. However, students who question the sources of knowledge and believe that knowledge is only constructed by the knower may have difficulty narrowing down multiple-choice questions because of their epistemic doubts (Bendixen & Rule, 2004).

Understanding why a student may produce an incorrect answer is valuable, and top-down examinations of misconceptions provide a reasonable explanation of what may be happening when a student answers an item incorrectly. When a misconception is held in memory, the individual is likely to feel confident that it is actually correct, especially when physical details can be incorporated into an explanation of the concept (Rozenblit & Keil, 2002; Sinatra et al., 2014). These details may affect the ontological structure of the explanation, creating a direct process explanation that is discrete, sequential, and independent rather than a more complex, emergent process explanation that relies on a deeper understanding of the collective action of the essential component parts of the concept (Chi, 2005; Rozenblit & Keil, 2002).

This understanding of top-down mechanisms does not describe what actually occurs when an incorrect answer is produced however. Indeed, incorrect answers can be quickly generated and reaction-like or they may be the result of slower, more analytical thought processes (Newman, Gibb, & Thompson, 2017). The behaviors of students during test taking may provide additional insight into how these top-down processes interact within the constraints of the testing environment.

Bottom-up cognition: Learner-specific and item-specific behaviors. The study of students' incorrect responses has focused mainly on the ways in which students may have misconceptions stored in memory. However, many students report knowing and understanding concepts, but they still produce incorrect answers in testing situations. Student behaviors on exams can be attributed to either the student (i.e., learner-specific factors), to the exam item (i.e., item-specific factors), or the interaction of student and item. Interestingly, many investigations have focused on learner-specific factors that drive testing behavior, but beyond describing the difficulty and discrimination of test items, item-specific factors have largely been unexplored.

Specifically, learner-specific investigations have centered on students' tendencies to change answers whereas item-specific investigations (where available) have centered on items that violate general item-writing guidelines.

Learner-specific behaviors. Two areas of student behaviors during testing have garnered attention in the research literature: answer change behaviors and response timing. Until recently, response time behaviors were largely ignored because they were impossible to measure on paper and pencil exams. However, with the growing use of computer software for testing purposes, response time behaviors are now being explored as a potential source of additional information on student performance.

Answer changing. Common wisdom advises students to stick with the first answer that comes to mind when answering multiple choice test items. This may not be the best course of action. Multiple studies (Bauer et al., 2007; Ferguson et al., 2002; Jeon et al., 2017; Stylianou-Georgiou & Papanastasiou, 2017) have found a benefit for students who reconsider answers. In fact, answer revision has been found to be beneficial 57-70 percent of the time (Stylianou-Georgiou & Papanastasiou, 2017). Obviously, answer revision is not a guarantee of additional points and sometimes students change answers from correct to incorrect, or incorrect to incorrect. However, as "The Gambler" advised, students would benefit from strategies associated with knowing when to "hold 'em" and knowing when to "fold 'em" (Rogers, 1978), in other words, students need to know when answer changing is most appropriate for them.

Traditionally, answer changing has been a relatively easy, unobtrusive method for examining students' testing behaviors. In most instances, answer changes were detected by examining student answer sheets for erasure marks, these were then used to determine whether the student changed an answer from right to wrong, wrong to right, or wrong to wrong

(Couchman et al., 2016; Jeon et al., 2017). Thus, a student's tendency to revise, and the skill that they display in their revisions are easily detectable.

Overall, students who change answers increase their exam score by two to three points for every point lost from revision (Benjamin et al., 1984). This benefit is greater for students who perform better on the exam, students in upper division classes, those who have additional time (or an instruction) to reconsider, and students who have a moderate level of confidence in their overall performance ability (Fischer et al., 2005; Geiger, 1996; Stylianou-Georgiou & Papanastasiou, 2017).

Jeon and colleagues (2017) described a three-node model for explaining answer change behavior on exams. In this model, the result of a students' initial answer (i.e., a correct or incorrect response) leads into one of two possible situations: if a correct answer is initially generated, the student could either keep the right answer or change it to wrong upon a revision; if an incorrect response is initially generated, then the student could change to a correct or the student could change to an incorrect response (or maintain the same incorrect response). In this model of answer changing, therefore, three latent traits of student testing behavior are present: the ability to generate initial correct answers, the ability to correctly change when the initial answer is incorrect, and the ability to make no change when the initial answer is correct (Jeon et al., 2017).

Three shortcomings exist when examining students' predilection for answer changing; the relatively low rate of answer changing among students (particularly as a student progresses to higher levels), the tendency of some students to change answers prior to marking an answer on the exam, and the inability to investigate behaviors beyond answer changing. Indeed, Couchman and colleagues (2016) identified the need to examine students' decisions to stick with an initial

answer in addition to their decisions to change answers. The ability to distinguish when it was most appropriate to stick with an answer versus when it was more appropriate to change an answer is dependent on the students ability to metacognitively monitor during exam performance; students with better metacognition (determined through measurements of confidence during the performance) were more likely to understand when to employ revision versus "sticking" (Couchman et al., 2016).

Response timing. More recently, researchers have begun to examine the length of time that students spend on questions and how response time may predict the generation of a correct or incorrect answer. Generally, response time has been used as an estimate of student motivation during low-stakes testing (Wise & Kong, 2005) or to examine the effects of time constraints on student performance (van der Linden, W. J., 2011). In these models, item response time is a factor for item construction governed by the speed-accuracy trade-off, where increased response time increases response accuracy. Students with low motivation during low-stakes tests will, therefore, answer questions more quickly and are predicted to generate more incorrect responses. Student motivation for answering is not considered a test construction issue for high-stakes testing because students are thought to approach these tests with a high value for correct answering (Wise & Kong, 2005).

Wang and Hanson (2005) developed an item response model incorporating response time for power tests. Modeling a student's response time on a power test presents some distinct advantages and disadvantages because these tests are designed to test a student's knowledge base. Unlike speeded tests, where items are designed to be easy enough to answer if the student has unlimited time to think about the questions, some students will provide incorrect responses on power test items regardless of the amount of time spent on the item. Incorporation of an "item

slowness" parameter in a model of item probability yielded a predictive equation that correlated well with item difficulty and discrimination indices (Wang & Hanson, 2005). However, in this model, response time and examinee parameters were assumed to be independent of one another. Thus, inferences about a person's ability level based on their response time to items cannot be made with this model. However, a hierarchical model proposed by van der Linden (2007) may allow for ability and response time comparisons.

The hierarchical framework for modeling speed and accuracy on test items (van der Linden, 2007) incorporates two levels of prediction. In the first level, item and person parameters are derived from data generated through testing (namely response accuracy and timing data). These parameters further predict the constructs of population and item domain. The main argument extended through van der Linden's (2007) model is that while speed and accuracy are independent of each other, item and person parameters interact to generate both accuracy and speed. Item parameters are identified as item difficulty and discrimination indices; however, "person parameters" were not specifically defined in the model (van der Linden, 2007). Therefore, a model examining both item difficulty and specific person parameters may have more explanatory power.

A feasible metric of a person parameter could be based on arguments from the "Top Down" perspective. Students who are more confident in the information they hold in memory should be able to produce answers faster than students who are unsure (Kruger & Dunning, 1999; Rozenblit & Keil, 2002). Thus, knowledge of a concept and/or a student's general background knowledge level may be important as a person parameter. This presents the possibility of a two-tiered response tree with the first tier representing correct and incorrect responses and the second tier representing the speed at which a response is generated (Figure 1).

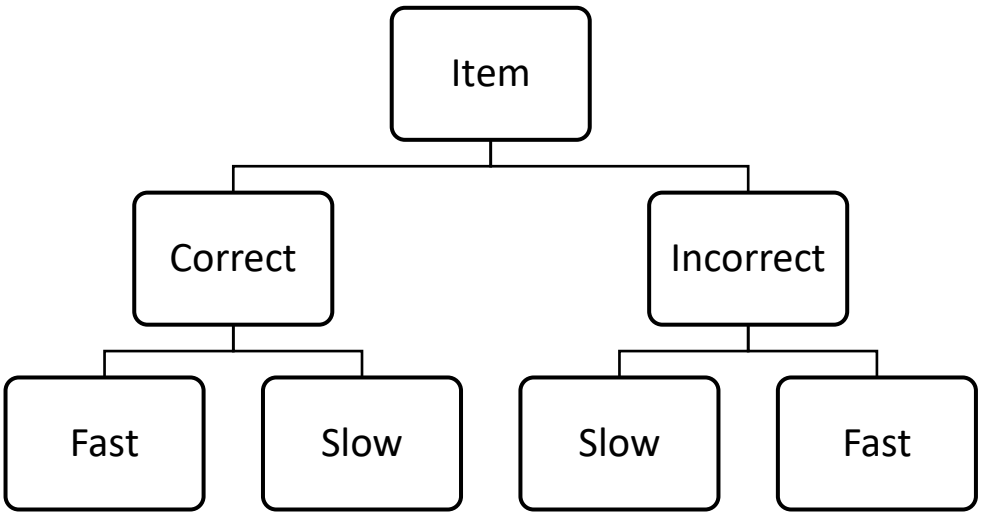

Figure 1: Tiered response and timing hierarchy for a given test item

As represented in Figure 1, correct and incorrect responses may be generated quickly or slowly, and the speed at which these responses are generated may indicate the underlying knowledge that a student has with respect to that concept. For example, a quickly generated incorrect response may indicate that a student has an ontological misconception regarding that concept, therefore the response that easily comes to mind is the response provided. However, quick responding may also be indicative of concept mastery when the student answers the question correctly. Interestingly, in this two-tiered representation of responses, answers that are generated slowly may indicate a level of knowledge somewhere between mastery and misconception, where the student needs to think through the reasoning behind the answer prior to marking a response. These students may have learned that their misconceptions are incorrect but

have not learned the concept to the point where the correct response is generated quickly. These students may be more likely to change an answer and more susceptible to item writing flaws.

Item-specific behaviors. Item-specific characteristics, such as stem length, distractor strength, item difficulty, and item discrimination may interact with students' ability to effectively employ test strategies. Jeon and colleagues (2017) noted that more difficult items yielded more answer changes, while easier items did not have many answer revisions. This provides some evidence that the interaction of student and item may be important when considering test construction.

Van der Linden (2007) also included item-specific behaviors in his hierarchical model of speed and accuracy, suggesting that more difficult items would require increased response times regardless of the ability level of the student. This assumes, however, that the students taking a test are equally motivated to provide correct responses to examination items. In addition, the model proposed by van der Linden (2007) assumes that individual students always approach items with the same answering strategy; while this may be true at the undergraduate level, this assumption may not hold true as students' knowledge level increases (Hess et al., 2015).

Item-specific characteristics have also been investigated with respect to item composition. Two common item composition errors are items that include nonsense distractors as answer choices and item stems that are difficult to interpret (Rush, Rankin, & White, 2016). These item writing issues have differential effects on item difficulty; nonsense distractors will reduce difficulty whereas a stem that is unclear increases item difficulty (Downing, 2005; Kiat, Ong, & Ganesan, 2018; Rush et al., 2016). When considered with the data on answer changing, these item flaws could put lower performing students at a distinct disadvantage (Downing, 2005; Jeon et al., 2017; Rush et al., 2016). For students in professional programs, item writing flaws

can impose additional stress in an already high-stakes environment, therefore putting struggling students at even more risk for poor performance. While the effects of item writing flaws on student performance have not necessarily been considered, utilization of test data on item timing and revisits may provide additional insight into how students behave when flaws are present.

Test Taking and Decision-Making in the Clinician

In addition to investigating test taking behaviors in undergraduate populations, several studies have examined test taking behaviors of physicians and residents (physicians completing their post-graduate training). For these populations, test taking behavior is used as a proxy for investigating decision-making strategies. Far from employing a single decision-making strategy, the data reveal that physicians are variable in their approach to decision-making, both between physicians (i.e., some physicians tend to make quick decisions while others are slower) and between instances (i.e., some instances cue fast decisions while others indicate a need to slow down). Indeed, several studies (Coderre, Wright, & McLaughlin, 2010; Eva & Regehr, 2007; McConnell et al., 2012; Moulton et al., 2010), indicate that physicians, particularly those with more experience, are acutely aware of when they need to slow down during decision-making and when they are comfortable relying on a quick, heuristic strategy for decision-making.

Examinations of test taking behaviors in groups of residents and physicians have revealed a complex relationship between response time, accuracy, and answer changing (Hess et al., 2015; Monteiro et al., 2015; G. Norman et al., 2014; G. Norman et al., 2016). These studies indicate that when physicians study an item for a longer time initially, they are more likely to generate an incorrect response, however when items were considered by difficulty level, an interaction between item difficulty and physician knowledge was discovered (Hess et al., 2015).

Hess and colleagues (2015) described the relationship between test taker ability (overall score), item difficulty, and response timing. These researchers found that on less difficult items, individuals with lower scores overall spent more time initially than those with higher scores. On more difficult items, increased response time was associated with more correct answers, regardless of how the individual performed on the rest of the exam. In addition, individuals benefitted from additional reflection on answer choices and this was inferred through an individual's likelihood to return to an item to change an answer (Hess et al., 2015). Similar to studies in undergraduate populations, answer changing benefitted the test taker more often than not, indicating that reconsidering answers is a beneficial strategy in testing.

When physicians are given specific instructions to answer quickly or answer slowly, no difference in diagnostic accuracy was discovered (Norman et al., 2014). In a similar experiment, Monteiro and colleagues (2015) directed experienced Emergency Room (ER) physicians and inexperienced ER residents to answer questions quickly or deliberatively. Individuals in this study were also randomly interrupted during test taking to simulate challenges in the ER environment. Again, directions were found to have little effect on diagnostic accuracy for either experienced physicians or residents. In addition, interruptions showed no interaction with accuracy or experience level (Monteiro et al., 2015). Since specifically directing test takers to perform a certain action (quick answer or deliberate answer) did not affect accuracy, suggesting that students approach testing from a "one size fits all" approach to test taking may do little to actually help a student develop better test taking skills. While it may not interfere with accuracy in the populations studied, directing individuals to always approach questions in a slow, methodical manner does slow test taking and may leave test takers with little time to consider more difficult questions that occur at the end of the test.

The studies examining residents and experienced physicians cite the fact that the populations studied may be too experienced to demonstrate differences in test taking and decision-making (Hess et al., 2015; Monteiro et al., 2015; G. Norman et al., 2014). Essentially, these researchers suggest that residents and practicing physicians are quite close in experience level, and that the manipulations studied in the experiments may yield more interesting results in less experienced populations. However, no studies have examined the combination of response time, item revisiting, and answer changing in medical students.

Emerging Patterns

A few patterns of test taking have been identified from the literature examining undergraduates and physicians. First, answer changing is more likely to be beneficial than harmful, especially when a test taker is more knowledgeable with respect to the tested subject. Second, response timing and knowledge display a complex relationship, whereby quickly answering questions can sometimes predict the likelihood of answering correctly; however slowly answering questions can also yield correct answers. While the test taking literature is inconclusive on the strategies that are most beneficial for students, investigations of decision-making have yielded a theoretical model that may help to frame the findings related to test taking. Pennycook, Fugelsang, and Koheler (2015) proposed a three-stage dual-processing model that incorporates both fast and slow decision-making and yields testable predictions for real world applications like test taking.

Theoretical Framework

Human reasoning is a complex and often fallible process. Indeed, a robust area of research exists that explores the ways in which humans make mistakes during the reasoning process (Evans & Stanovich, 2013; Kahneman, 2011; Pennycook et al., 2015). As originally

proposed by Tversky and Kahneman (1974), dual-process theory relies on two systems to explain the decision-making process. System one is conceptualized as a fast, heuristic decision maker that relies on emotional salience, environmental context, and experiential cues to reach a decision (Evans & Stanovich, 2013; Kahneman, 2011). System two is conceptualized as a slow, analytical decision maker that processes the logical relationships associated with the decision (Kahneman, 2011). The switch from system one to system two processing is thought to progress one of two ways: through direct intervention by system two when conflict is detected (Evans & Stanovich, 2013) or through parallel processing of the initial stimulus by both systems (Handley & Trippas, 2015).

Pennycook and colleagues (2015) provided a model for the progression of thinking from system one into system two (see Figure 2). In their model, system one is instead called type one

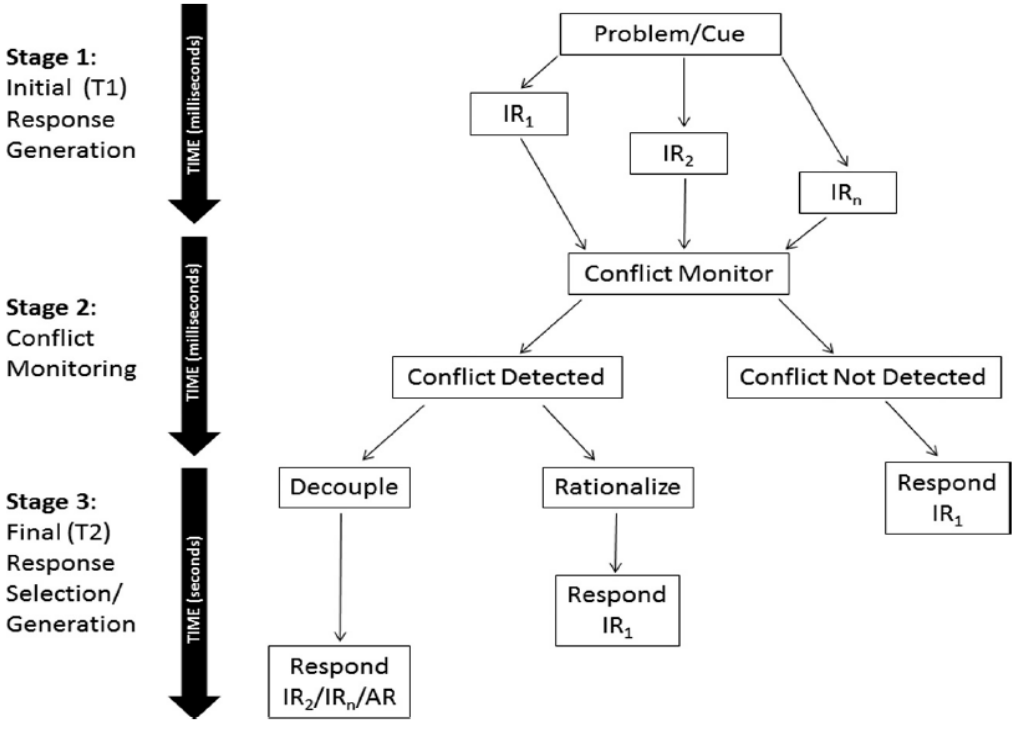

Figure 2: The three-stage, dual process model for thinking. From: Pennycook, G., Fugelsang, J. A., & Koehler, D. J. (2015).

(T1) thinking, whereas system two is converted to type two (T2) thinking; this terminology will be adopted from this point forward. The three-stage dual processing model cites conflict detection as a key rationale for the progression of thinking from T1 to T2 processes (Pennycook et al., 2015). In their model, T1 thinking is cued by a stimulus (stage 1), which could lead to a conflict detection (stage 2), thus invoking T2 thought processes for conflict resolution (stage 3).

In a test-taking situation, the "problem/cue" represented in the model would be the stem for the item under consideration. The initial response (IR_1, IR_2, IR_n) would be the initial responses that come to mind when the student reads the item. The conflict monitor is enacted when multiple initial responses are generated for the item and there is no intuitively "best" response among the responses generated. If a conflict is detected among the intuitive responses, additional thought is required to either rationalize the first response generated or decouple from that response to support a different response. Thus, the student could answer the item quickly with an intuitive response or answer the item slowly with an intuitive or analytic response.

Type 1 Processing

Type one thought processes generate fast, intuitive responses, mainly through heuristics based on environment, emotion, and/or experience (Gigerenzer & Brighton, 2009; Kahneman, 2011). Early research on T1 decisions focused on inherent biases in T1 thinking that predispose individuals to illogical thought patterns (Tversky & Kahneman, 1974); more recent research has explored how T1 thinking can sometimes be biased, but may also be accurate (Bago & De Neys, 2017; Gigerenzer, 2008; Pennycook et al., 2015). These research paradigms have used expertise frameworks (Gigerenzer & Brighton, 2009; Klein et al., 1986) and base-rate problems (Bago &

De Neys, 2017; Pennycook, Trippas, Handley, & Thompson, 2014) to explore accuracy and bias in T1 thought processes.

Expertise Frameworks. Development of domain-specific expertise could be a useful method for investigating the accuracy of T1 processes and the possibility of conflicts between parallel T1 outputs. Two common frameworks for conceptualizing expertise are Fast and Frugal Heuristics (FFH; Gigerenzer & Brighton, 2009) and Naturalistic Decision-Making (NDM; Klein, 2015). While there are differences between how these two frameworks conceptualize expert decision-making, they are not mutually exclusive.

The FFH framework proposes that as individuals gain expertise, they add heuristic rules to their "adaptive toolbox" (Gigerenzer & Brighton, 2009). Several heuristics, or "rules of thumb" have been proposed to explain expertise effects in decision-making. The "recognition" heuristic is used when one option is better known than the other option; this could be employed by novices and experts alike as recognizable information often leads to accurate decisions (Marewski, Schooler, & Gigerenzer, 2010). The "fluency" heuristic predicts that information that is quickly and easily recalled will be used for decision-making (Marewski et al., 2010). A third heuristic, "take-the-best", describes a method of ignoring cues to speed decision-making (Gigerenzer & Brighton, 2009). This heuristic also improves with expertise because experts are better at knowing which stimuli are high-yield for the decision at hand, attending to those stimuli, and using them as inputs for the T1 process (Gigerenzer, 2008).

The NDM framework conceptualizes expertise as the ability to recognize patterns in the environment that support a certain course of action (Klien, 2015). As individuals progress from novice to expert, they are exposed to a variety of patterns in their chosen field, some of which are common, some uncommon (Klein et al., 1986). With repetition, common patterns are

remembered and these patterns become easier to recognize thus priming specific response patterns (Klein et al., 1986). Uncommon patterns are also easy to detect because of peculiarities that do not match the patterns stored in memory; the more expertise one has, the easier it will be to spot patterns that are both common and uncommon (Klein, 2015).

The FFH and NDM frameworks can be used to explain the parallel streams hypothesis of T1 thinking (Martin & Sloman, 2013). If an expert has developed multiple, trustworthy heuristics, then parallel outputs will be generated from T1 thinking. For example, if all three heuristics (recognition, fluency, and take-the-best) are applied to diagnostic decision-making, a doctor is likely to generate three ideas for a diagnosis (Wegwarth, Gaissmaier, & Gigerenzer, 2009). Conversely, through the NDM framework, a pattern of symptoms may indicate a certain disease, but certain symptoms may belong to multiple disease patterns that the doctor has stored in memory. Depending on the individual's domain-specific experience with that set of symptoms, the FFH or NDM frameworks could generate congruent (i.e., all three diagnoses match) or incongruent T1 outputs. If the diagnoses are congruent, the doctor assigns a most likely diagnosis and begins a treatment plan (Wegwarth et al., 2009). If the output of the T1 process is incongruent, however, it is predicted that further T2 processing would be enacted to arrive at a diagnosis (Pennycook et al., 2015).

Expertise has been a valuable model for exploring accurate decisions that are made with fast heuristics, but T1 decisions, whether correct or incorrect, are also characterized by bias. In order to explore how biased, incorrect, and often difficult to overcome decisions are generated by T1, researchers have used extreme base-rate problems to manipulate the aforementioned expertise effects from T1 processing.

Base-Rate Problems. Base-rate problems typically use a suggested population with an imbalanced subset of the population (e.g., 1000 people were surveyed, 995 nurses, 5 doctors; Pennycook et al., 2014). This population sample is then contrasted with a diagnostic description of one individual (e.g., John is 34 years old, lives in a beautiful home in a posh suburb, is well-spoken, and interested in politics). The descriptions are followed by a question relating to the single individual's status (e.g., Is John a doctor or a nurse?). Typically, participants in base-rate experiments will neglect the numerical description of the sample in favor of the diagnostic description. In the example provided, participants will typically state that John is a doctor even though it is more likely that John is a nurse because of the number of nurses included in the sample (Kahneman, 2011; Pennycook et al., 2014; Tversky & Kahneman, 1974). Tversky and Kahneman (1974) termed this heuristic as "base-rate neglect" and demonstrated that this heuristic is robust and resistant to change unless additional analytic (T2) resources are incorporated into problem solving.

Examination of participants' response time and the addition of response pressure (such as speeded response or cognitive load) has shown that individuals incorporate some processing of base-rates during T1 responding (Bago & De Neys, 2017; Pennycook et al., 2014). Pennycook and colleagues (2014) demonstrated that extreme base-rate problems, such as the above example, resulted in increased response time even when participants provide a response in line with the stereotype description. This increase in response time is thought to correspond to a conflict between T1 outputs as T2 processing was occupied with a cognitive load task. Bago and De Neys (2017) pursued the idea of conflicting T1 outputs further and have shown that answer change depends on the individual's perception of conflict between T1 outputs; individuals who rate their initial answers with lower confidence are more aware of conflict between T1 outputs

and are more likely to change their answer with further processing time (Bago & De Neys, 2017).

Type 2 Processing

Type two thought processes are slow, analytical responses that are based upon explicit knowledge (Kahneman, 2011). The initiation of T2 thought processes were once thought to guarantee "correct" response generation (Tversky & Kahneman, 1974). However, intuitive or illogical slow responses that incorporate T2 processing have been observed. The strength, or the number of times a T1 response has generated a positive outcome, is predictive of the T2 processing that will occur; strong T1 responses that have worked in the past will generate rationalizations, while weaker T1 responses that are being learned may generate cognitive decoupling (Pennycook et al., 2015).

Rationalization. Once a decision has been passed on to analytic processing, an individual still can provide an intuitive response, this is accomplished through the process of rationalization. Rationalization focuses on justifying the T1 response through elaboration; effortful processing is induced by verification of the intuitive T1 response rather than falsification of that response (Pennycook et al., 2015).

Rationalization has been observed within the study of base-rate neglect (Newman et al., 2017) and syllogistic reasoning (Handley & Trippas, 2015). With respect to base-rates, participants were more likely to provide responses supported by beliefs based on the stereotype of the target individual when given additional time to reconsider their answer (Newman et al., 2017). In syllogistic reasoning studies, participants were more likely to spend additional time thinking about why a syllogism is believable rather than why a syllogism is logical (Handley &

Trippas, 2015). These studies indicate that T2 processing can invoke the individual's belief framework to support conclusions derived from T1 processes.

While rationalization may not seem harmful considering the experiments that have described the relationship between processing time and belief, investigations of conspiracy theory also point to a relationship between time allowed for thought and increased reliance on belief, particularly when factual information is absent (Lewandowsky, Oberauer, & Gignac, 2013; Lewandowsky et al., 2015). In addition, rationalization has been observed as an underlying factor in situations where individuals rate their knowledge level higher than what it may be (Kruger & Dunning, 1999). Rationalizations have also been found to increase one's belief that a self-generated explanation sufficiently describes physical and natural phenomena (Rozenblit & Keil, 2002).

Cognitive Decoupling. Considering all the ways that thinking can produce an incorrect response, it is sometimes surprising that we can generate analytic responses that differ from intuitive responses. However, cognitive decoupling is the response most typically associated with T2 processes (Evans & Stanovich, 2013). Cognitive decoupling can be defined as the process by which an intuitive response is inhibited and overridden by a competing intuition, a separate intuitive response generated by T1 processing, or an alternative response generated by T2 processing (Pennycook et al., 2015). Unlike rationalization, the focus of decoupling is falsification of the intuitive response or weighing two responses against each other to determine which is most appropriate (Evans & Stanovich, 2013).

Cognitive decoupling has been investigated in experts and novices in a wide variety of experiential situations. Pretz (2008) examined social decision-making in first through fourth year college students, finding that fourth year college students were more likely to generate logical

decisions after participating in analytic processing. In addition, Gervais (2015) found that providing individuals who were predisposed to analytic thinking with additional time to think enabled those individuals to override previously held creationist beliefs and endorse evolution. Finally, a robust line of research examining performance of physicians in a wide variety of complex situations has described the benefits of "knowing when to slow down" in order to reduce diagnostic or treatment errors (Chang et al., 2016; Eva & Regehr, 2007; Hruska, Hecker et al., 2016; McConnell et al., 2012; Moulton et al., 2010; R. Norman et al., 2017).

Distinguishing between T1 and T2 responses allows for investigation of how fast and slow responses are generated and how prior experience can affect both types of response. While some responses will be generated solely through T1 processing, the switch to T2 processing is also a source of research inquiry. Namely, researchers have questioned the nature of the T1 outputs, the "decision maker" regarding T2 processing, and how individual differences may affect the likelihood of T2 engagement.

Moderating Change Between the Systems

Two theoretical approaches to T1/T2 switching are present in the literature: the default interventionist (Evans, 2006; Kahneman & Frederick, 2005) and the parallel processing approach (Martin & Sloman, 2013; Sloman, 1996). Both approaches derive from the need to explain how response processing may progress from T1 to T2, the generation and interpretation of error signals, and the outcome of process switching.

Default Interventionist. Tversky and Kahneman (1974) originally proposed a monitoring mechanism of T2 that processes T1 outputs for acceptability, fluency, or relevance with current goals (relevancy principle; Evans, 2006). Low level analytic processing is thought to start as soon as the decision-making process begins. When a T1 response is generated, it is

generally accepted unless a rejection of the response is deemed necessary (satisficing principle; Evans, 2006). T2 monitoring of T1 responses is generally casual; many T1 responses are expressed even though they are not appropriate because the T2 monitor is "lazy" (Kahneman & Frederick, 2005).

Errors generated through heuristic decision-making mechanisms in T1 result in full engagement of T2 analytic processing. T2 processing is not immune to the biases passed along from T1 processing; response fluency, confidence, and emotional valence are passed along to T2 processing from T1 to "aid" in the decision-making process (Darlow & Sloman, 2010; Evans, 2006). These ties to T1 processing are thought to speed analytic processing; however, remnants of heuristic responses may also increase the likelihood that a T1 response will be rationalized through the analytic process rather than decoupled (Kahneman & Frederick, 2005).

Parallel Processing. Several responses may be generated by a stimulus simultaneously in T1 processing. Some of these responses may rely solely on T1 heuristics while others may rely on T2 logical rule-based processing (Martin & Sloman, 2013; Sloman, 1996). The parallel processing approach posits that T1 and T2 processing of a stimulus begin at the same time. If a response is generated quickly through T1 processing, then it is accepted and expressed. If the T1 an T2 responses are generated in relatively similar times, then further processing is necessary through T2 (Martin & Sloman, 2013; Sloman, 1996). The difference between T1 and T2 processing (e.g., time to process, confidence, emotional valence) will determine the size and importance of the error signal (Darlow & Sloman, 2010). Again, T2 processing could be biased by remnants from the T1 response, but parallel processing hypothesizes little interaction between responses generated by T1 onto T2 (Martin & Sloman, 2013; Sloman, 1996).

Cognitive Reflection Test. One of the most common methods for measuring an individual's tendency towards T1 or T2 responding, and the ability to change between T1 and T2 responses, is the Cognitive Reflection Test (Frederick, 2005a). The original three-item instrument includes open-ended questions that are likely to generate an incorrect T1 response. This test has been recently expanded to seven items (Toplak, West, & Stanovich, 2014) and both the seven- and three-item tests are reliable for predicting intuitive and analytic responding (Frederick, 2005a; Pennycook, Cheyne, Koehler, & Fugelsang, 2016; Toplak et al., 2014).

An example of an item included on the CRT is: A bat and a ball together cost $1.10, a bat costs $1.00 more than a ball. How much does a ball cost? Generally, 65% of individuals provide the intuitive answer to this question (Pennycook et al., 2016). The intuitive answer is "10 cents" however the correct answer, usually arrived at through deliberation, is "5 cents" (example calculation: if the ball costs 10 cents, the bat would cost $1.10 cents based on the criteria set forth in the problem, thus adding up to $1.20; if the ball costs 5 cents, the bat would cost $1.05, adding up to $1.10). When considering the generation of this response in terms of the three-stage, dual-process model, the intuitive, T1, answer is generated quickly and, for some individuals, this answer is the only answer generated. Other individuals may detect a conflict between the T1 answer and the correct answer similar to the conflict described when individuals detect errors in base-rate problems (Bago & De Neys, 2017; de Neys, Rossi, & Houde, 2013; Pennycook et al., 2016). Again, strength of the T1 response is indicative of the individual's ability to override the incorrect intuitive answer through cognitive decoupling or the tendency to rationalize the T1 response (Pennycook et al., 2015).

The patterns of incorrect answers on the CRT may provide an explanation of a student's tendencies on their first visit to an item and reconsideration of answers to multiple choice items.

Students who generate a single answer quickly may detect no conflict in their thought processes with respect to the question at hand. Again, this quickly generated answer may be attributable to a learned response and may be correct or the answer may be incorrect and therefore based on a biased response pattern (Pennycook et al., 2015; Stylianou-Georgiou & Papanastasiou, 2017). Further reconsideration, evident in the form of a longer first-view time or additional time spent on an item after an initial answer is chosen reflect the use of T2 processing (Couchman et al., 2016; Pennycook et al., 2015).

CRT and thinking disposition. Thinking disposition is the tendency of an individual to rely on T1 or T2 processes for decision making. The original three-item CRT assigns thinking disposition by majority answer; if the individual chose two or more intuitive answers, the assigned thinking disposition would be "Intuitive" (Frederick, 2005a). Direct measurement of the tendency to use T1 or T2 processes is advantageous because individuals can over or underestimate their preferences for thinking, especially those who rely on T1 processes as these processes typically occur subconsciously (Pennycook et al., 2015).

Szaszi, Szollosi, Palfi, and Aczel (2017) examined individual reaction times when answering CRT items correctly and incorrectly. Interestingly, they found four groups of responders quick and correct, slow and correct, slow and incorrect, and quick and incorrect. These results support the idea that some individuals may be able to generate T1 responses that match the analytical response pattern, possibly based on their experiences with numbers and calculations, while others rely on heuristics that typically lead to incorrect response patterns (Szaszi et al., 2017). This research supports the three-stage, dual-process model proposed by Pennycook and colleagues (2015) because the slow responding participants generated both correct (cognitive decoupling) and incorrect (rationalization) responses (Szaszi et al., 2017). This

may also indicate that the amount of time a student spends considering an item could provide important insight into the student's thinking patterns and knowledge base.

CRT, metacognition, and confidence. Many factors affect an individual's confidence in a generated answer (Chi, 2005; Rozenblit & Keil, 2002). Confidence, it seems, has an inverse relationship with both knowledge and tendency to rely on T1 processes for decision-making (Kruger & Dunning, 1999; Pennycook et al., 2017). When participants were asked to rate their performance on the CRT in addition to assessing their need for cognition using the *Need for Cognition* scale (Pacini & Epstein, 1999); Pennycook and colleagues (2017) found that those who were most likely to provide the incorrect intuitive response to CRT items were more likely to have high confidence in their answers and also most likely to overstate their use of analytic thinking. Furthermore, as actual CRT score increased, confidence in that score decreased, such that the individuals who answered the most items correctly underestimated their performance by a factor of 1.1 (Pennycook et al., 2017).

Based on the work of Rozenblit and Keil (2002), individuals who generate shallow explanations of concepts generally rely on more intuitive thought processes. Because these thought processes are less accessible to conscious thought, they are less likely to generate conflict and more likely to instill feelings of confidence (Pennycook et al., 2017). Monitoring confidence and the ability to use confidence judgements as a threshold for decision making have been shown to interact with performance on the CRT. Individuals with high general knowledge, low answer confidence, but high confidence thresholds are best able to override their intuitive responses and provide the correct response on the CRT (Jackson et al., 2016).

Following this line of logic, physicians who do not fully understand the underlying basic sciences may be more likely to confidently follow their intuitively generated diagnoses and less

likely to know when it may be appropriate to think more analytically about a diagnosis (Coderre et al., 2010; Eva et al., 2010; Moulton et al., 2010). Conversely, physicians who are more analytical may be hampered in their decision making because of their tendency to over-analyze information due to lower levels of confidence (Hess et al., 2015; Reach, 2014). However, while dual-process theory has been proposed as a model for diagnostic thinking, no research has explicitly related performance on theoretical measures like the CRT and tests of a physician's knowledge base, such as multiple-choice test items.

Summary

Throughout the existing literature on test taking, several models and theories attempt to explain how students approach the decision-making process when answering questions on multiple-choice exams. Top-down theories of misconceptions provide insight into why students may produce incorrect answers, but provide little information about how these answers were produced. Bottom-up descriptions of student behaviors provide interesting descriptions of what students are doing during exams but provide little connection to the process of decision-making. Empirical studies with physicians and residents shed light on expert decision-making in clinicians but do little to describe the development of this process. Decision-making theories have explored tendencies to use intuitive or analytical thought processes for decision-making but lack application to decision-making outside of experimental situation.

Ideally, a model for decision-making during test taking would begin with an observable behavior such as the amount of time a student takes to consider an item when it is first encountered. This provides an initial measurement of the student's decision making with respect to the concept presented in the test item. Extending the hierarchical model proposed by van der Linden (2017), person-level parameters could be specified as the student's thinking disposition,

current knowledge, and/or prior knowledge. Item difficulty and correct/incorrect responding provide item-level and student-by-item parameters. As detailed by the three-stage dual process model for decision-making, these factors could influence a student's initial consideration of an item and may relate to the student's ability to interpret and adjust item answering strategies.

The Present Study

The present study employed the three-stage, dual process model as a framework for test-taking behaviors during a high-stakes, multiple-choice exam in the second year of medical school. A relationship was proposed between the first-time an item was encountered and the response to that item. Based on the predictions of the three-stage, dual process model, this relationship should be negative; such that items answered incorrectly should have longer first-view time.

Person-level factors (current knowledge, prior knowledge, and thinking disposition) and item difficulty were hypothesized to moderate the relationship between timing and response. Current knowledge and prior knowledge were expected to strengthen the relationship between response and first-view time; individuals who scored higher on the exam, or those with higher first-year GPA, were expected to spend more time considering items answered incorrectly and less time on items answered correct. Knowledge level (either previous knowledge or current knowledge) should enable students to generate more intuitive correct answers, but should also cue activation of the conflict monitor and T2 thought processes when an intuitive correct answer is not immediately generated.

Thinking disposition was expected to weaken the relationship between first-view time and response, such that if an individual provided more correct answers on the CRT, that person would display less difference in first-view time between correct and incorrect answers. A student

who is able to answer more CRT items with the correct (analytical) answer is more likely to recruit T2 resources when challenged by a question. Therefore, these students should spend more time on questions overall, but should have less difference in first-view time between their correct and incorrect answers because they are more likely to use T2 thinking for all answers.

In addition to these person-level factors, item difficulty, an item-level factor, has been shown to increase response time. Therefore, a multilevel mixed model incorporating answer-level (response), item-level (item difficulty) and person-level (thinking disposition, current and prior knowledge) factors was proposed to explain how a student interacts with test items.

Chapter 3: Methods

Multiple choice exams are often used to assess learning; some of these exams are used to grade students' course specific knowledge while other exams are used to determine the student's career path after graduation. Generally, all exams taken by medical students are termed "high-stakes" because the results of these exams weigh on a student's grade point average and class rank. These measurements of student ability are considered by residency directors to determine post-graduate medical training.

High-stakes exams can provide valuable insight into both the behaviors that students employ when making decisions in high stress environments and the students' underlying knowledge base. However, traditional techniques used to measure students' behaviors when approaching exam items, such as "talk aloud protocols," can interfere with students' ability to perform and are not suitable for measurement of students while they are taking an actual exam. The advent of computer-based testing offers an opportunity to examine students' behaviors as they are taking an exam and, when paired with theoretical input regarding Type 1 (intuitive) processing and Type 2 (analytical) processing, a deeper understanding of a student's knowledge base may be possible. Indeed, this may provide the ability to understand a student's answer beyond "right or wrong".

Research Design

To investigate thinking disposition, testing behaviors, and performance on a high-stakes multiple choice exam, several measurements must occur without interfering with the overall performance of the student. Prior investigations of thinking disposition have occurred only in laboratory settings and investigations of test performance have not been able to measure students' decision making during the exam. Therefore, this study was unique in that thinking

disposition was measured in conjunction with the student's performance on the current exam and the student's historical performance in medical school.

Thinking Disposition

The present study employed a multiple-choice version of the Cognitive Reflection Test (Primi et al., 2016; Sirota & Juanchich, 2018; Toplak et al., 2013) to examine traditional thinking dispositions among second-year medical students. This study used the CRT to establish a student's tendency for intuitive or analytical thinking then related that disposition to behaviors during high-stakes testing. The four-choice format used by Sirota and Juanchich (2018) has shown construct equivalence to the original open-ended format described by Frederick (2005b). The CRT appeared as an optional survey at the end of a subject-based exam for medical students in the first semester of their second year of medical school.

Current and prior knowledge

Aggregate data from the responses found in the snapshot file were used to generate a measure of the student's current knowledge (exam score- total number of correct answers on exam items divided by the total number of items on the exam). Participants in the study granted access to their first-year grade point average (GPA- average performance on all first-year courses on a 4.0 scale) therefore establishing a measure of prior knowledge. These measures, in conjunction with the thinking disposition, constituted the person-level factors for the multilevel mixed model.

Testing behavior

Examplify (ExamSoft Inc., 2017) records a "snapshot" each time a student enters an item, when a student selects an answer, when a student leaves an item, and if the student returns to the

item. These exam snapshots are always recorded through the software and were initially intended to provide a safety measure against catastrophic software failure. The exam snapshot file was reduced using the protocol described by Miller and Lindquist (2016) to yield the amount of time that a student spends initially viewing an item (first-view time) and the answer the student provided for the item (response).

Participant Selection and Data Collection

The study took place at a small osteopathic medical school in the Southwestern United States. The experimental methods for this study were reviewed and approved by the Institutional Review Board at the University of Nevada, Las Vegas. Participants were able to opt-in to data collection via an informed consent form presented at the end of a subject-based exam during the first semester of their second-year of medical school.

Statistical analysis

A three-level multilevel mixed model incorporating person-level and item-level factors was specified for data analysis. This statistical procedure was chosen because each student in the participant pool answered each question in the item pool. A traditional ANOVA would therefore be confounded by the level at which it was conducted (e.g., if the ANOVA were conducted at the student-level, there would be data for 16,632 test items). Therefore it is important to build the statistical analysis with considerations of the multiple levels apparent in the data collected.

The outcome variable (dependent variable) for this study was the amount of time a student spent when first encountering an item (first-view time). The student's performance on the exam, GPA, and thinking disposition were used as person-level (level three) factors. The difficulty of the item was used as the item-level (level two) factor. Response (correct/incorrect) was used as the student-by-item (level one) factor.

Summary

A three-level mixed model of student-by-item (level one) in item-level (level two) in person-level (level three) was needed to understand student behavior during exams. The model described here only investigated one aspect of behavior, first-time view of an item, however other behavioral aspects such as number of visits or total time spent on an item could be substituted as the dependent variable. At level one and two, response correctness and item difficulty have been shown to affect exam behavior (Hess et al., 2015; van der Linden, W. J., 2009). The addition of person-level (level three) factors as moderators of the relationships between response/first-time view and difficulty/response/first-view time should provide insight into how students moderate their behavior during high-stakes testing.

Chapter 4: Results

The research questions posed for this study are: Is there a relationship between testing behavior and the response ultimately generated by the student? If there is a relationship between testing behavior and response, what is the impact of factors such as item difficulty, student knowledge, and/or thinking disposition on this relationship? To address these questions, a multilevel mixed model was specified with the initial time a student spent on an item as the dependent variable, the response (incorrect/correct) as an answer-level (level one) factor, the difficulty of an item as an item-level (level two) factor, and current knowledge (exam score), historical knowledge (first year GPA), and thinking disposition (score on CRT) as person-level (level three) factors.

Description

One-hundred eight students from the second-year medical class at a small osteopathic school elected to participate in the study. The second-year cohort is comprised of 129 students (40% female, average age = 23.4 ± 2.7 years), the study population reflected the demographics of the cohort at large. The target exam for the study was the 154-item cardiovascular system final exam. The exam length and the number of students who volunteered for the study resulted in performance data for 16,632 observations for analysis.

Students provided informed consent by clicking “yes” on an informed consent item presented at the end of a computer-based course exam. The items for the Cognitive Reflection Test (CRT) were inserted after the informed consent. Only data from students who answered yes to the informed consent item were included in the study, even though some students elected to answer the CRT items without providing consent. Students were made aware that the informed

consent and survey would appear at the end of their exam and the research protocol was approved by the Institutional Review Board at the University of Nevada Las Vegas.

The CRT was scored by awarding one point per correct answer. Scores ranged from 0 (all incorrect) to 3 (all correct). For statistical analysis, CRT responses were coded by the number of responses correct (0,1,2, and 3). All statistical analyses using the CRT scoring were referenced to the highest level of the indicator (i.e., 3 or "all analytic/correct answers"). Therefore, the categories of thinking disposition were "intuitive", "majority intuitive", "majority analytic", and "analytic". The CRT results yielded 24 students who answered none of the questions correctly (intuitive). Twenty-seven students provided one correct answer on the CRT (majority intuitive). Twenty-three students answered two of the CRT questions correctly (majority analytic). Finally, 34 students provided correct answers to all three questions on the CRT instrument (analytic).

Students who participated in the study allowed access to their de-identified testing data, including exam score and exam snapshots. They also allowed the researcher access to their GPA, calculated at the end of the first year in medical school. Average exam score for the study participants was 78.72% (SD = 6.53). Average first year GPA for the study participants was 3.26 (SD = 0.26).

Prior to conducting statistical testing, item difficulty, GPA, and exam score were grand-mean centered. Centering these factors was important because there was not a zero-value for them in the data set. Therefore, item difficulty, GPA, and exam score were centered on their respective grand means because interpreting the intercept at zero would not make sense (Heck, 2014). Results of the CRT were not centered because there were individuals within the sample who scored zero on the instrument and a zero score on the CRT is indicative of an intuitive thinking disposition.

Responses to exam items were coded one (1) for a correct answer and zero (0) for an incorrect answer. Item response constituted the student-by-item level for statistical analysis. Incorrect response served as the baseline (comparison) for all statistical analyses.

Null Model

The first step in the multilevel mixed modeling procedure was to generate a null model. The null model answers the question: Does first-time view vary across questions and students? The grand mean for the length of first-time view per student per item was 54.10 seconds. The intraclass correlation (ICC) was calculated by taking the ratio of the variance between groups at a specific level divided by the total variance. This calculation compares the variance within the grouping structure for a particular level to the overall variance in the data set. Stated differently, the ICC is the expected correlation of two data points within the same group; at higher values, the ICC indicates more homogeneity within groups than between groups. The ICC for level one was 0.5942, which indicates that the variance occurring between groups at level one (student-by-item) in first-view time was 59.42%. At level two the ICC was 0.3302; thus, the variance between questions at the item level was 33.02%. Finally, the ICC at level three was 0.076, indicating 7.6% variance between students in first-time view.

Intercepts varied significantly across students (Wald $Z = 6.959$, $p < 0.05$), items (Wald $Z = 8.596$, $p < 0.05$) and student-by-item (Wald $Z = 90.474$, $p < 0.05$). This variation, in combination with the ICC value suggest there is sufficient variance at each of the levels to continue with the multilevel analysis (Heck, 2014).

Random intercept model

The multilevel mixed linear modeling procedure calls for including the factors from all three levels as predictors in the equation for the random intercept model. The estimated intercept for the random intercept model is 65.44 seconds, this is the average first-view time on an incorrectly answered item while holding factors at levels two and three constant. The effects of the factors at each level are summarized in Table 1.

Entering the factors for each level resulted in a 17.66% decrease in variance at the item level, a 4.28% decrease in variance at the person level, and a 2.7% decrease in variance at the answer level. The variance at all three levels was still significant (level one Wald $Z = 90.47$, $p < 0.05$; level two Wald $Z = 6.79$, $p < 0.05$; level three Wald $Z = 8.54$, $p < 0.05$), these values indicate that there was still significant variability to be explained.

Table 1

Main effects of factors at each level on first-view (random intercept model)

	Estimate	Std. Error	Sig. (p)
Level 1			
Response	-15.15	0.71	< 0.05
Level 2			
Difficulty	-0.40	0.09	< 0.05
Level 3			
GPA	1.98	1.40	0.16
Exam Score	-0.25	0.06	< 0.05
Intuitive	3.10	0.75	< 0.05
Maj. Intuitive	2.84	0.72	< 0.05
Maj. Analytic	-3.85	0.75	< 0.05
Analytic[a]	-	-	-

Note[a]: Analytic thinking disposition was the reference category for all analyses including thinking disposition

Random slope and intercept model

The random intercept and slope model was used to determine whether the level one and two factors (response and item difficulty) contributed to the observed differences on first-time view across students.

The primary interest in generating this model was the estimates of variance components for the random slopes (Heck, 2014). The variance in the slope of response/first-time view was significant at level two (Wald $Z = -3.78$, $p < 0.05$), which indicates that the slope of first-time view based on response was different depending on the item difficulty. Similarly, the variance was significant (Wald $Z = -5.26$, $p < 0.05$) when the slope of response/first-time view was free to vary at the level three. Finally, the variance of the relationship between item difficulty and first-time view was significant at level three as well (Wald $Z = -4.54$, $p < 0.05$). These significant variance components indicate that there is the possibility of cross-level interactions within the model (Heck, 2014)

Cross-level interactions model

The questions for this stage of model building were: Are there combinations of factors across levels that moderate the relationship between response and first-time view? The specific interactions investigated at level one were: level two (item-level difficulty) on the relationship between response and first-time view, level three (person-level thinking disposition and exam score) on the relationship between response and first-view time. At level two, the interaction of interest was between person-level factors (thinking disposition and exam score) and the relationship between item difficulty and first-time view. Finally, a three-way interaction was specified such that the level three factors moderate the impact of difficulty on the relationship between response and first-time view. GPA was not specified in the cross-level analysis because it failed to reach significance in the original random slope model.

Results for the main effects of factors included in this model are summarized in Table 2. Results for interactions included in this model are summarized in Table 3.

Table 2

Main effects of factors at each level on first-time view (full model)

	Estimate	Std. Error	Sig. (p)
Level 1			
Response	-15.59	1.95	< 0.05
Level 2			
Difficulty	-0.21	0.10	< 0.05
Level 3			
Exam Score	-0.15	0.20	0.44
Intuitive	7.20	3.59	< 0.05
Maj. Intuitive	4.55	3.42	< 0.05
Maj. Analytic	-0.63	3.60	0.86

Note: Analytic thinking disposition was the reference category for all analyses including thinking disposition

Table 3

Cross-level effects on first-time view (full model)

	Estimate	Std. Error	Sig. (p)
Level 2 on Level 1			
Difficulty*Response	-0.39	0.08	< 0.05
Level 3 on Level 2			
Score*Difficulty	-0.01	0.01	0.13
Intuitive*Difficulty	0.10	0.70	0.15
Maj. Intuitive*Difficulty	-0.07	0.07	0.33
Maj. Analytic*Difficulty	0.07	0.07	0.30
Level 3 on Level 1			
Score*Response	-0.09	0.11	0.46
Intuitive*Response	-5.74	2.07	< 0.05
Maj. Intuitive*Response	-2.90	1.95	0.14
Maj. Analytic*Response	-4.95	2.06	< 0.05
Three-way interactions			
Score*Diff.*Response	0.01	0.01	0.13
Intuitive*Diff.*Response	-0.004	0.09	0.96
Maj. Intuitive*Diff.*Response	0.13	0.09	0.12
Maj. Analytic*Diff.*Response	0.09	0.09	0.34

Note: Analytic thinking disposition was the reference category for all analyses including thinking disposition

Since exam score failed to reach significance in the cross-level model, a second cross-level model was estimated with thinking disposition as the sole level 3 factor. In addition, the interaction between level three and level two, and the three-way interactions were also eliminated from the final model due to the lack of significant results found in the initial cross-level interaction model. Removing these factors from the analysis led to a more parsimonious model with response at level one, item difficulty at level two, and thinking disposition at level three. This model had two specified cross-level interactions: first-view time by item difficulty and response, and first-view time by thinking disposition and response.

The final, parsimonious multilevel mixed model yielded an intercept of 66.20 seconds, this reflects the prediction for the time to answer an item of average difficulty incorrectly for a student with analytic thinking disposition. Main effects remained the same across levels, however the interaction between thinking disposition and response became significant for all but one of the thinking disposition categories (Table 4).

Table 4

Interaction effects for the parsimonious multilevel mixed model

	Estimate	Std. Error	Sig. (p)
Level 2 on Level 1			
Difficulty*Response	-0.34	0.06	< 0.05
Level 3 on Level 1			
Intuitive*Response	-4.98	1.84	< 0.05
Maj. Intuitive*Response	-3.34	1.77	0.60
Maj. Analytic*Response	-4.43	1.86	< 0.05

Note: Analytic thinking disposition was the reference category for all analyses including thinking disposition

The influence of response on first-time view (level 1) is negative, indicating that students spend about 15 seconds less on an item when generating correct answers than incorrect answers. At level 2, students spend .21 seconds less for every 1% increase in difficulty index (increasing difficulty index indicates easier items). To visualize these results, a graph was created using first-view times of the most difficult items (< 50% of students responded correctly), first-view time for items around the average (< 80% of students responded correct), and first-view time for the highest performing items (< 95% of students answered correct; figure 3).

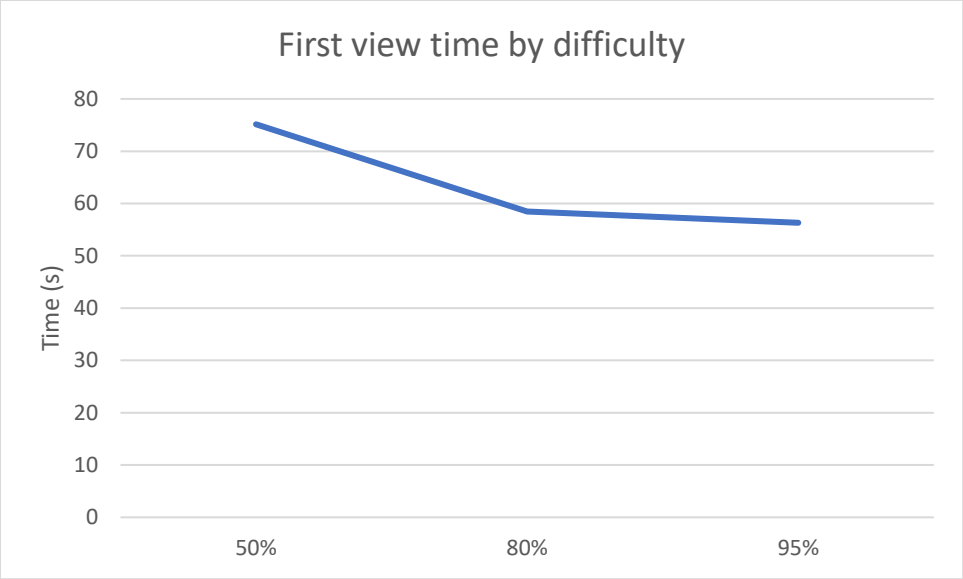

Figure 3: Aggregate first-view time by item difficulty level

Finally, at level 3, students with more intuitive thinking dispositions spend more time looking at items initially than students with more analytic thinking dispositions (7.2 seconds for students who answered all CRT items incorrectly, 4.5 seconds for students who answered one CRT item correctly). To visualize this relationship, first-view time was aggregated by thinking disposition categories, then graphed versus the predicted values (figure 4). As can be seen, the multilevel model slightly overestimates the amount of time spent by each thinking disposition category, but the overall relationship between the thinking dispositions is similar.

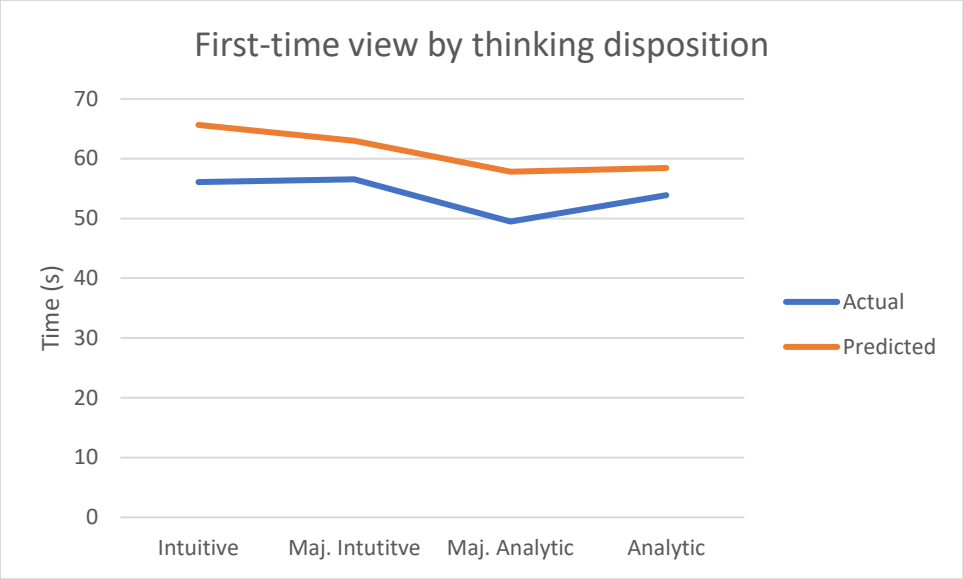

Figure 4: Actual versus predicted first-view time across thinking dispositions

The interactions suggest that both item difficulty and thinking disposition moderate the response/first-time view relationship. The effect of item difficulty on the response/first-time view relationship is -0.34 seconds, which indicates that as items get easier students spend even less time on correct answers. To illustrate this relationship, an aggregate first-time view was created for the five most difficult items (~50% of students answering correctly), five items around the average difficulty level (~80% of students answering correctly), and five least difficult items (~95% of students answering correctly) then these averages were graphed (figure 5).

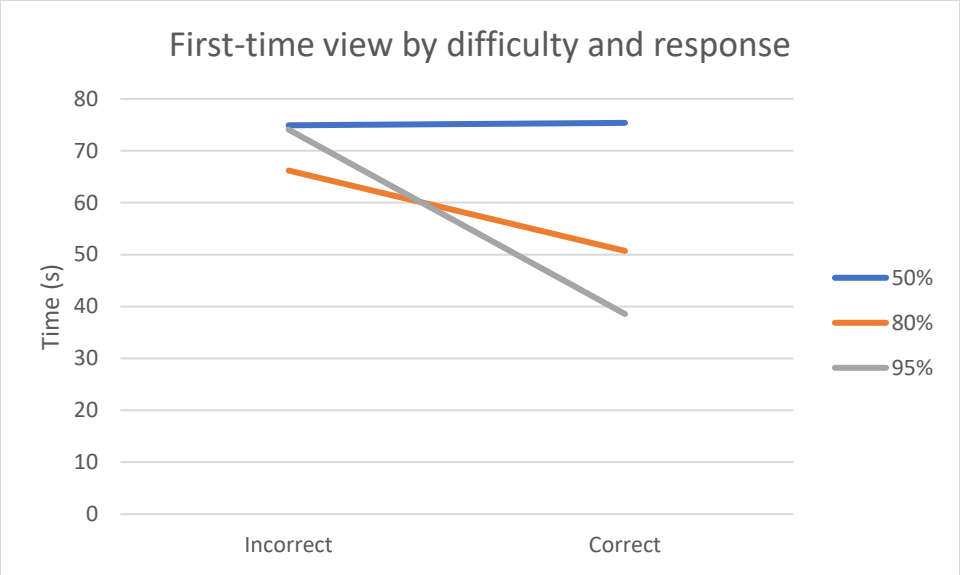

Figure 5: Aggregate first-view responses by difficulty and response type

The interaction between first-view time, response, and thinking disposition is negative overall, indicating that students in each of the thinking disposition categories spend less time on correct than incorrect answers. Again, actual first-view times for each of the thinking disposition categories on correct and incorrect responses were aggregated and graphed (figure 6).

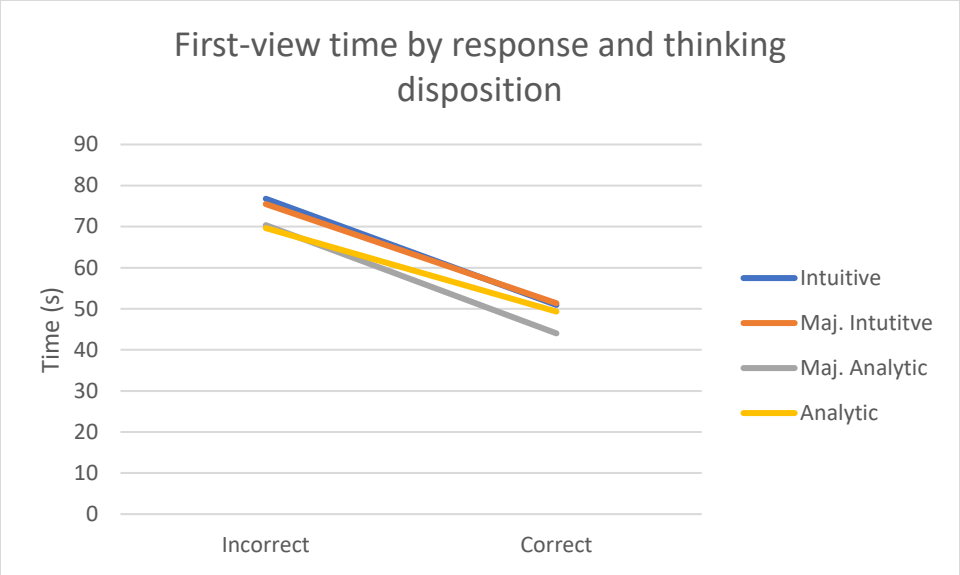

Figure 6: Aggregate first-view time for thinking disposition by response

The variance statistics for the final model suggest that the cross-level interactions account for the majority of the variance initially seen in first-view timing due to response (Wald Z = 1.884, p = 0.60). The variance for response reduced by 49.82% over the variance initially estimated when response was specified as a random parameter in the random intercept and slope model. Variance at the item-level was still significant (Wald Z = 8.16, p < 0.05), as was variance at the person-level (Wald Z = 6.16, p < 0.05), indicating there are additional factors beyond those included in the model that affect first-view time at these levels.

Summary

A multilevel mixed model was specified to address four questions related to student behaviors on exams. The null model tested whether there were differences between items and between students on timing when they are first exposed to an item. The null model described significant variance across students (Wald Z = 6.959, p < 0.05), items (Wald Z = 8.596, p <

0.05), and answers (Wald Z = 90.474, $p < 0.05$). The intraclass correlation coefficient (ICC) was also significant for level one (.5942), level two (.3302), and level three (.076). The variance components and significant clustering effects indicated by the ICC calculations indicated enough separation between groups to continue with model construction.

Second, response type (correct/incorrect) was added as a level one factor to determine whether responding correctly or incorrectly to an item affected the amount of time a student spent on first-view. This relationship was significant and negative (-15.15 s, $p < 0.05$) indicating that students spend roughly 15 seconds less when they first view items to which they ultimately provided correct answers. After establishing the main effect of response on first-view time, the slope of response was specified as a random variable in the random intercept and slope model. The slope of the first-time view response relationship was found to vary at different levels of item difficulty (Wald Z = -3.78, $p < 0.05$). This relationship also significantly varied based on the factors at level three (Wald Z = -5.26, $p < 0.05$).

Item difficulty was also added as an item-level (level 2) factor in the second stage of multilevel modeling. Item difficulty significantly influenced first-view time (-0.40 s, $p < 0.05$) which indicates that for every 1% increase in item difficulty from average (easier items), students spent almost a half second less on first time item view. Significant variance was found in the first-view time and difficulty relationship (Wald Z = -4.54, $p < 0.05$) when this relationship was specified as a random variable at level 3.

At level three, three person-level factors were added to the model to determine whether current knowledge (exam score), prior knowledge (first year GPA), and/or thinking disposition (CRT score) affected the amount of time a student spends viewing an item for the first time. Main effects were found for exam score (-0.25 s, $p < 0.05$), therefore students spend about a

quarter second less on first time view per one-point increase over the average exam score. Thinking disposition for each category was also significant (intuitive: 3.10 s, $p < 0.05$; majority intuitive: 2.84 s, $p < 0.05$; majority analytic: -3.85 s, $p < 0.05$). All values for the main effects of thinking disposition used analytic thinking disposition as a reference category. Finally, GPA was not a significant predictor of first-view time (1.98 s, $p = 0.16$), therefore GPA was not used in any other stage of the multilevel modeling process.

The third and final step in the multilevel mixed modeling process was to examine cross-level interactions between the factors. In the cross-level model, the main effect of exam score was no longer significant (-0.15 s, $p = 0.44$), indicating that the original main effect detected was possibly due to item difficulty or the interaction between item difficulty and response. Based on this, exam score was eliminated as a person-level factor for the final, parsimonious multilevel model. The interaction between thinking disposition and item difficulty was also non-significant (omnibus $F = 2.07$, $p = 0.10$). In addition, the three-way interaction between thinking disposition at the person-level, item difficulty at the item-level, and response at the answer-level factors was not significant (omnibus $F = 1.104$, $p = 0.346$), demonstrating that the effects of item difficulty on the relationship between first-view time and response are separate from the effects of thinking disposition. Therefore, the three-way interaction and the interaction between thinking disposition and difficulty were dropped from the final model.

In the final, parsimonious multilevel model, the effect of item difficulty on the relationship between first-time view and response was significant (-0.39 s, $p < 0.05$), suggesting that as items become easier there is greater difference between the time students spend on incorrect and correct answers. The effects of thinking disposition on response were significant at two levels (intuitive*response: -5.74 s, $p < 0.05$; majority analytic*response: -4.95 s, $p < 0.05$),

which suggests that thinking disposition category had a differential effect on the relationship between first-view time and response. These results are discussed further and related to the theoretical framework in the discussion chapter.

Chapter 5: Discussion

The main research question addressed in this study was is there a relationship between testing behavior and the response ultimately generated by the student? If there is a relationship between testing behavior and response, what is the impact of factors such as item difficulty, student knowledge, and/or thinking disposition? Based on the results of the final multilevel mixed model, there is a relationship between the amount of time a student spends looking at an item for the first time and the response that the student ultimately generates. A student's current knowledge and their past knowledge did not moderate the relationship, however thinking disposition and item difficulty significantly affected both first-view time and the relationship between response and first-view.

Response, difficulty, and thinking disposition

The three-stage, dual process model for decision making (Pennycook et al., 2015) predicts that students should spend more time thinking when multiple responses are generated through T1 (fast, intuitive) thought processes. Conflict between these responses triggers additional analysis by T2 (slow, analytic) processing in order to generate an answer. Analysis by T2 is thought to proceed through either rationalization of incorrect responses or cognitive decoupling to generate correct responses. This hypothesis was supported by the data in this study. Students spent about 15 seconds longer on first-view for items answered incorrectly than items answered correctly. This could mean that students, when considering an item for the first time, detected a conflict among multiple responses to the item and proceeded with T2 processing.

Item difficulty provided additional insight into how students behaved when encountering items for the first time. When initially considering a difficult item, students spent about the same amount of time on incorrect and correct answering, but on easy items the difference in first-time view between correct and incorrect responses was 44 seconds faster for correct responses. This supports the hypothesis forwarded by Pennycook and colleagues (2015). When considered in conjunction with the three-stage, dual processing model for decision-making, this may indicate that students, when faced with difficult test items, spend equal amounts of time rationalizing incorrect responses and decoupling to generate correct responses.

As test items become easier, students spend less time on correct answers. Essentially the responses to easier items are more intuitive and therefore result in less need to incorporate additional processing prior to registering a response. Interestingly, for incorrectly answered items, students spent slightly less time on responses for items at the medium difficulty (~80% correct response) level than difficult or easy items (74.9 s for difficult, 66.2 s at medium difficulty, 74.1 s for easy). This may be an indication of over-rationalizing incorrect answers on these items; students may be convinced that the answer "could not be that easy" and therefore choose a more complex response to the item. Indeed, this behavior has been anecdotally observed in the medical student population (Burns, 2006) and may be a result of the speed at which students are expected to learn complex scientific and medical information (D'Eon, Kosmas, & Macmillan, 2007)

The impact of thinking disposition on the relationship between response and first-view time is also predicted by the three-stage, dual process model for decision making (Pennycook et al., 2015). Students who chose more analytic, correct answers on the CRT (two or three correct

answers) spent less time on items answered incorrectly. However, students with the most analytic answers on the CRT were more likely to spend the same amount of time answering items correctly as the students who were more intuitive. While this amounted to a small difference on the slope between incorrect and correct responses, the result was still significant. Seconds are important in the decision-making realm; stage 1 (T1 response generation) and stage 2 (conflict monitoring) occur in the millisecond range while stage 3 (T2 response generation) occurs in the second range. Therefore, a significant one second change in decision making does have explanatory benefit.

Based on the predictions of the three-stage, dual process model, students who answer more items correctly on the CRT could be more flexible thinkers than those who answer incorrectly on the CRT, and this could result in less difference on timing between correct and incorrect answers. The difference between the students who scored high on the CRT and the students who scored low on the CRT may be the direct result of the rationalization/decoupling process. A student who scores high on the CRT may be more likely to engage T2 processes in decision making, thus slowing the generation of correct answers because these answers are cognitively decoupled from incorrect responses generated by T1 thinking. Alternately, a student who scores low on the CRT may be more likely to rely on T1 processes to generate answers. This results in a fast answer, but when this student detects a conflict between T1 responses, he or she may be less motivated to engage in T2 processing or may engage in rationalization to support an incorrect response (Pennycook et al., 2017).

Current and prior knowledge

The non-significant impact of both current and prior knowledge on first-view time was surprising and suggests a more complex relationship between knowledge and decision-making than was modeled here. The three-stage dual process model suggests a strong effect of knowledge on response timing. Based on the predictions of the model, students with the most knowledge (high exam score or high GPA) should generate fast, correct answers because the information has been learned to an "intuitive" responding level. Likewise, students who score lower on the exam, or have a lower GPA would be expected to take more time when generating answers because they have detected a conflict in the answers generated by T1 thought processes and need to use T2 thinking to either decouple or rationalize incorrect responses (Pennycook et al., 2015).

Why did knowledge have little to no impact on first-time view? Prior research on decision-making suggests five explanations for the non-significant results related to knowledge. First, students who scored low on the exam may have provided intuitive incorrect answers to items. If this were the case, the relationship between knowledge and first-view time may be quadratic in nature; both high knowledge and low knowledge individuals would generate answers quickly while students with moderate amounts of knowledge take longer to consider item content. Indeed, this result is suggested by the three-stage, dual process model for decision-making and is a target for future research.

Second, the significant impact of thinking disposition on the relationship between response and first-view may mask the effect of knowledge. Since students who score higher on the CRT spend less time on items overall, perhaps these students have more practice with T2

thinking and can therefore proceed through the rationalization/cognitive decoupling process faster than those who choose more intuitive answers on the CRT.

Third, item difficulty was not included as an item-level factor in the multilevel mixed linear model. The difficulty index for an item is generated by determining how many students answered the item correctly divided by the total number of students who provided an answer (Allen & Yen, 1979) Items that are more difficult may take longer for a student to answer than items that are easier, resulting in a knowledge by difficulty interaction on the relationship between response and first-view. However, students may also recognize that an item is more difficult and elect to skip the item and return to it later or put down a "best guess" for the item without returning, further complicating the relationship between knowledge, response, and first-view timing.

Fourth, the length of an item stem should impact first-view time. The exam used in this study included multiple clinical vignette style questions requiring a student to read through a paragraph of patient information to generate a diagnosis. Additional time may be required if the item required that the student synthesize the information and provide a treatment or pathogen for the diagnosed disease. In this situation, while the student might perform well on the exam, first-view time would be longer because of the length of the item stem. Conversely, some shorter items may also require longer first-view times because they do not provide enough information for the student to quickly generate an answer. For these items, increased first-view time would be related to the student searching for information pertaining to the item in memory rather than the rationalization/cognitive decoupling process.

Finally, item responses were only scored as correct or incorrect on the exam, the actual answer chosen by the student may provide additional insight into how knowledge relates to responding and first-view time. If a student has an incomplete knowledge of the subject tested, an answer that appears correct may be chosen based the student's intuitive or superficial understanding of the concept being tested. This answer would therefore be generated quickly and require little additional T2 processing. Grading the answer chosen based on the point-biserial for each answer option could result in a more significant relationship between knowledge, response, and first-view timing.

In a timed testing environment, incorrect answering could prove costly; spending additional time on first-view may result in items not seen at the end of an exam. Using the three-stage dual process model as a predictor of decision-making, incorrect answers most likely take more time to generate because students are rationalizing their incorrect responses. Correct responses are generated more quickly because students have learned information while studying for the exam. Differences between students who score high on the CRT and those who score low on the CRT may be attributable to differences in experience with or motivation for engaging T2 thought processes. Finally, the non-significant effect of knowledge on first-view, response, or the relationship between responding and first-view may be confounded by several different factors.

Implications

Based on the literature related to decision-making by clinicians, an important characteristic of good physicians is the ability to "know when to slow down" (Moulton et al., 2010). This flexible approach to decision making demonstrated by the students with more

analytical answers on the CRT was evident in the more similar treatment of correct and incorrect answers produced on exam items.

In general, all of the students in the study spent less time initially considering correct versus incorrect answers. However, when thinking disposition was added to this, students with higher CRT scores spent less time overall on the items, but had a shallower slope between incorrect and correct responses. These results imply that students who are more willing to use analytical (T2) thought processes may be more likely to slow down when confronted with challenging information. This predisposition to slow down when selecting correct answers may be beneficial because it enables the clinician to incorporate new evidence into the decision-making process (Coderre et al., 2010).

From an educational standpoint, the relationship between response and first-view time is important. The time associated with generating incorrect answers may be an ideal target for implementation of context-based, test-taking strategies. If students were provided with timing feedback after an exam, they might be able to recall the decision-making processes used to generate the answer. This might help them to identify more productive thought patterns that promote cognitive decoupling over rationalization. If timing feedback were available during an exam, students might be able to use this to selectively revisit and rethink items that required more time to answer.

Finally, students may also want to examine their individual tendencies with regard to response timing. Despite the absence of a significant relationship between knowledge, response, and first-view described by multilevel mixed model specified here, some students may detect patterns in their own timing when they feel particularly confident or less-than-confident in their

subject-based knowledge (Rangel, Möller, Sitter, Stibane, & Strzelczyk, 2017). Building conscious awareness of decision timing may also help students who score lower on the CRT to identify the necessary cues in their own thoughts that signal the need to slow down, ultimately helping students to become better decision makers.

Limitations

Neither the student's current knowledge nor their prior knowledge appeared to affect behavior on the exam items. This seems contrary to the predictions of the three-stage, dual process model because students with higher knowledge levels should generate more intuitive correct answers, or should generate more competing responses with T1 thought processes that need to be resolved with T2 thought processes. Significant variability between- and within-students remained in the final model including current and prior knowledge, indicating that there may be other factors at the person- or item-level that have an impact on the relationship between knowledge, behavior, and response.

At the item-level, adding additional factors to the analysis such as discrimination index or item length could result in a more nuanced understanding of how students interact with exam items. An additional consideration for the answer-level may be the actual answer selected on the exam, which would indicate whether the student was selecting the most reasonable distractor or an answer choice that is more of an intuitive misconception. If significant, these factors could then be used to provide additional information to item writers regarding student performance. In addition, this information could be used by exam administrators to predict the time necessary for a high-stakes exam.

At the person-level, additional factors suggested by a recent study on individual differences in decision making competence may provide additional explanatory power (Talukdar, Román, Operskalski, Zwilling, & Barbey, 2018). Monitoring decision-making by fMRI while study participants responded to items on the Adult Decision Making Competence (A-DMC) battery showed significant contributions to decision-making by executive functions, social and emotional factors, and somatosensory and perceptual processes. While the A-DMC is significantly longer than the CRT used in this study, the A-DMC could be provided to students outside of a testing environment then the results could be related to decision-making during the exam. This would provide a more in-depth description of person-level decision-making factors that may result in a more significant explanatory model.

One of the strengths of this research, measuring student behavior during an actual exam, also presents its own limitations as well. Because the exam used in this research was designed to test student's clinical knowledge of the cardiovascular system, the items included on the exam were not specifically directed at eliciting intuitive or analytic responses. Faculty who wrote the exam items may have included some intuitive distractor choices, but there were no specific instructions to include these answer choices when items were considered for the exam. A future direction of research could be to specifically include some items with intuitive distractors and measure student responses to those items compared to similar items without the intuitive distractor. This may provide more insight into how item writing may direct students toward or away from specific answer choices.

The items included on the exam were also specific to medical students and the practice of medicine. While previous studies have been more specific to medical practice (Eva & Regehr,

2007; Hess et al., 2015; Monteiro et al., 2015; G. Norman et al., 2014), this study used students who are still learning the practice of medicine. These students were specifically selected because they are learning the process of diagnostic decision making. Studies of more general, high-stakes exams, such as the GRE, MCAT, or LSAT that specifically target analytic thinking in item design, may reveal different patterns of answering among students with more analytic or intuitive thinking dispositions.

Conclusions

Decision-making is a complex process that has long drawn the attention of researchers. Tversky and Kahneman first proposed a dual stage decision-making model in 1974, resulting in multiple research studies investigating intuitive and analytic decision-making, and the rules that govern the switch between the two types of thinking. Recently, Pennycook and colleagues (2015) proposed a three-stage dual process model to govern decision making. This model provided testable hypotheses regarding decision-making. Importantly, quickly generated responses (i.e., intuitive answers) can be correctly generated when an individual has learned a response pathway, and slowly generated responses (i.e., analytical) may indicate that an individual is in the process of resolving a conflict between multiple possible answers. The results of this study support this conclusion; students generally responded faster to items that they answered correctly and slower on items that they answered incorrectly. This result indicates that students were slowing decision-making when faced with information that they did not know well. However, this research does not indicate whether students were consciously aware of their behaviors, which could be a target for further research on student behavior during exams.

Earlier studies of exam behaviors have primarily focused on answer changing by students. These types of investigations provided useful information, such as a tendency of students to benefit from answer changing, especially for more advanced students (Fischer et al., 2005; Geiger, 1996; Stylianou-Georgiou & Papanastasiou, 2017). However, past studies have been unable to directly measure student behaviors such as response timing due to the use of paper-and-pencil based tests. This study used computer-based testing that recorded exam activities through snapshot data. These data, when combined with the results of the CRT, provided useful information regarding a person-level factor that affects decision-making during testing.

Interestingly, the students in the study were evenly divided on the CRT scores; 24 students answered no questions correct, 27 answered one question correct, 23 answered two questions correct, and 34 answered all of the questions correct. This result illustrates the point made at the start of this research – medical students are not a homogeneous group of stellar students. In fact, these students come from diverse backgrounds and have learned many strategies that have led them to where they are today. For some, intuitive decision making is a smart strategy that enables them to use keyword searching during exam taking to reduce cognitive load and maximize responding while minimizing effort. For others, taking time to think through an item enables them to connect pieces of information and develop deeper understanding of the diagnostic process. While there are drawbacks to both strategies, there are also benefits to viewing exams and testing in these ways as well. Future research could explore this avenue further through conducting interviews with students through the medical school process. This may identify key points in time where student thinking about exams and exam items shift, furthering understanding of how students interact with the exams they are taking.

The first key finding in this research was that students who provided more correct answers on the CRT generated faster responses during testing, but also had a smaller differential between first-view of items with correct versus incorrect responses. This insight, that students who are able to generate analytical responses on the CRT are more flexible with their exam taking behaviors, is important because it provides a link between decision-making theory and behaviors expressed in non-experimental settings. This finding is also important because flexible decision-making is a highly valued characteristic in clinicians. The participants in this study, second-year medical students, are on the path to becoming clinicians and will soon need this ability to effectively practice medicine.

The second key finding was that when students are responding to difficult items, they spend equal amounts of time considering correct and incorrect answers. As items become easier, first-view time decreases for items answered correctly, but stays the same for items answered incorrectly. These relationships were predicted by the three-stage, dual process model proposed by Pennycook and colleagues (2015), and ultimately demonstrate that, when challenged, students do slow the decision-making process in order to think through the choice they are making. Again, the ability to slow down the decision-making process is important for clinicians, but it could be argued that slowing down is meaningless if it only leads to rationalization. Students need to identify when they are rationalizing answers and switch strategies to enable cognitive decoupling. By developing more conscious exam taking behaviors, students may become better at "knowing when to hold 'em and when to fold 'em."

Appendix 1: Mulitlevel Modeling Procedure

In order to specify a three-level multilevel mixed model that incorporates item-level and person-level factors, several steps must be completed in data analysis (Aguinis, Gottfredson, & Culpepper, 2013; Heck, 2014). First, a null model was considered to describe differences between answers, questions, and participants in the study. Second, a random intercept, fixed slope model was identified to detect differences between factors at level one, two, and three on the dependent variable. Third, a random intercept, random slope model was specified to determine the variance in the relationship between the factors at levels one and two and the dependent variable. Finally, estimation of a cross-level interaction model allows for understanding whether a particular factor at any level, or if combination of factors across levels, were able to explain at least part of the variance in the relationship response and first-time view. These steps correspond to four questions related to the current research problem:

1. Null model: Does first-time view vary across students, items, or student-by-item?
2. Random intercept model: Do thinking disposition, current performance, GPA (level three factors) explain differences in first-time view? Does question difficulty (level two factor) explain differences in first-time view? Does response (level one factor) explain differences in first-time view?
3. Random slope and intercept model: Does the item-level factor of difficulty explain variation in the relationship between response and first-time view? Do the person-level factors of current knowledge, prior knowledge, and/or thinking disposition explain the relationship between difficulty or response and first-time view?
4. Cross-level interactions model: Are there combinations of factors across levels that moderate the relationship between response and first-time view?

Null model

The null model was used to determine differences in first-view time between questions and between students. Rejection of the null model would mean that questions differ in the amount of time students spend initially considering them. The level 1 equation for the null model is:

$$Y_{ijk} = \pi_{0jk} + \sum_{p=1}^{P} \pi_{Pjk}\, a_{Pijk} + \varepsilon_{ijk}$$

This equation reflects that the first-time view for the student-by-item interaction *i* on question *j* for student *k* is equal to the intercept (π_{0jk}) plus level one coefficient (π_{Pjk}) multiplied by the the response (level 1 predictor = a_{Pijk}) plus the level one residual (ε_{ijk})

Similarly, at level 2 the null model equation is:

$$\pi_{pjk} = \beta_{p0k} + \sum_{q=1}^{Q_p} \beta_{pqk}\, X_{qjk} + r_{pjk}$$

Where β_{p0k} represents the intercept for question *k* in modeling the question effect on first-time view, X_{qjk} represents the level 2 predictor (question difficulty), β_{pqk} represents the corresponding level 2 coefficient, and r_{pjk} is the level 2 random effect.

At level 3 the null model equation is:

$$\beta_{pqk} = \gamma_{pq0} + \sum_{s=1}^{S_{pq}} \gamma_{pqs}\, W_{sk} + u_{pqk}$$

Where γ_{pq0} represents the intercept in modeling the person-level effect on first-time view, W_{sk} represents the level 3 predictors (thinking disposition, current knowledge, and prior knowledge), γ_{pqs} represents the corresponding level 3 coefficient, and u_{pqk} is the level 3 random effect.

In order to complete this step, the intraclass correlation (ICC) was calculated at all three levels. This specified the amount of clustering associated with the grouping structure for any level of the data set (Heck, 2014). The ICC was calculated using the equation:

$$ICC = \sigma^2_{leveli} / [\sigma^2_{level\,1} + \sigma^2_{level\,2} + \sigma^2_{level\,3}]$$

Where $\sigma^2_{level_i}$ equals the variance at a particular level. Generally, a cut-off of 0.05 is recommended for the ICC (Heck, 2014). For this study, a significant value for the ICC would reflect statistically significant variability in first-view time for the level at which it was calculated.

Random intercept model

The random intercept model describes the relationship between the at the three levels and first-time view (Aguinis et al., 2013; Heck, 2014). The equation for calculating the random intercept model at level one is as follows:

$$Y_{ijk} = \pi_{0jk} + \pi_{1jk} response_{ijk} + \varepsilon_{ijk}$$

The level one random intercept model equation predicts first-time view for answer *i* to question *j* for student *k* (Heck, 2014).

At level two, the measure of question difficulty was added. This measure was group mean centered for the question so that the resulting intercept reflected the first-time view for an item of average difficulty. The equation for the level two random intercept model is:

$$\pi_{0jk} = \beta_{00k} + \beta_{01k} difficulty_{ijk} + r_{0jk}$$

$$\pi_{1jk} = \beta_{10k}$$

Finally, at level 3, the factors of thinking disposition, current knowledge (exam score), and prior knowledge (GPA) were added as predictors. Current and prior knowledge were grand mean centered yielding a predicted first-time view for a student who is analytical with an average exam score and average GPA.

$$\beta_{00k} = \gamma_{100} + \gamma_{001} TDcat_k + \gamma_{002} score_k + \gamma_{003} GPA_k + u_{00k}$$

$$\beta_{10k} = \gamma_{100}$$

$$\beta_{01k} = \gamma_{010}$$

Combining the equations together yields nine parameters to estimate; six fixed effects (the grand mean intercept, γ_{000}, plus the effects for each factor at each level), two random effects (u_{00k} and r_{0jk}), and the level one residual (ε_{ijk}). The level three intercepts are therefore adjusted for response at level one and difficulty at level two. The complete equation is:

$$\beta_{00k} = \gamma_{000} + \gamma_{100} response_{ijk} + \gamma_{010} difficulty_{jk} + \gamma_{001} TDcat_k + \gamma_{002} score_k + \gamma_{003} GPA_k + u_{00k} + r_{0jk} + \varepsilon_{ijk}$$

Random intercept and slope model

The random intercept and slope model determines if the level one and two factors (response and question difficulty) contribute to the observed differences on first-time view across students. The question posed here was: does student response or item difficulty explain

differences between questions on first-time view between students? The equation to indicate a random slope for the response effect is:

$$\pi_{1jk} = \beta_{10k} + r_{1jk}$$

The equation to indicate a random slope for the question difficulty effect is:

$$\beta_{01k} = \gamma_{010} + u_{01k}$$

Substituting these equations into the combined equation adding the random slope parameters to level 3:

$$\beta_{00k} = \gamma_{000} + \gamma_{100} response_{ijk} + \gamma_{010} difficulty_{jk} + \gamma_{001} TDcat_k + \gamma_{002} score_k + \gamma_{003} GPA_k + r_{1jk} response_{jk} + u_{01k} difficulty_k + u_{00k} + r_{0jk} + \varepsilon_{ijk}$$

This equation yielded an estimate (intercept) first-view time for an item of average difficulty answered incorrectly by a student with an analytical thinking disposition, average GPA, and average exam score (Heck, 2014).

Cross-level interactions model

Question difficulty was predicted to impact the response/first-time view relationship. Additionally, level three factors (thinking disposition, current knowledge, prior knowledge) should impact the difficulty/first-time view relationship, the response/first-time view relationship. In addition, a three-way interaction was specified between student, question, and answer (level three, two, and one, respectively). To investigate the influence of difficulty on the response/first-time view relationship, the level two equation is:

$$\pi_{0jk} = \beta_{00k} + \beta_{01k} difficulty_{jk} + \beta_{12k} (difficulty * response_i)_{jk} + r_{0jk}$$

The level three equation is:

$$\beta_{00k} = \gamma_{100} + \gamma_{001}TDcat_k + \gamma_{002}score_k + \gamma_{003}GPA_k + \gamma_{011}\left(TDCat * difficulty_j\right)_k$$
$$+ \gamma_{012}(score * difficulty_j)_k + \gamma_{013}(GPA * difficulty_j)_k$$
$$+ \gamma_{101}(TDcat * response_i)_k + \gamma_{102}(score * response_i)_k$$
$$+ \gamma_{103}(GPA * response_i)_k + \gamma_{111}(TDcat * difficulty_j * response_i)_k$$
$$+ \gamma_{112}(score * difficulty_j * response_i)_k$$
$$+ \gamma_{113}(GPA * difficulty_j * response_i)_k + u_{00k}$$

Substituting the cross-level terms into the combined equation results in the final model equation:

$$\beta_{00k} = \gamma_{000} + \gamma_{100} + \gamma_{001}TDcat_k + \gamma_{002}score_k + \gamma_{003}GPA_k + \gamma_{011}\left(TDCat * difficulty_j\right)_k$$
$$+ \gamma_{012}(score * difficulty_j)_k + \gamma_{013}(GPA * difficulty_j)_k$$
$$+ \gamma_{101}(TDcat * response_i)_k + \gamma_{102}(score * response_i)_k$$
$$+ \gamma_{103}(GPA * response_i)_k + \gamma_{111}(TDcat * difficulty_j * response_i)_k$$
$$+ \gamma_{112}(score * difficulty_j * response_i)_k$$
$$+ \gamma_{113}(GPA * difficulty_j * response_i)_k + u_{00k} + r_{0jk} + \varepsilon_{ijk}$$

References

Aguinis, H., Gottfredson, R. K., & Culpepper, S. A. (2013). Best-practice recommendations for estimating cross-level interaction effects using multilevel modeling. *Journal of Management, 39*(6), 1490-1528. doi:10.1177/0149206313478188

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Long Grove, IL: Waveland Press.

Bago, B., & De Neys, W. (2017). Fast logic?: Examining the time course assumption of dual process theory. *Cognition, 158*, 90-109. doi:10.1016/j.cognition.2016.10.014

Bauer, D., Kopp, V., & Fischer, M. R. (2007). Answer changing in multiple choice assessment change that answer when in doubt--and spread the word. *BMC Medical Education, 7*(1), 28. doi:10.1186/1472-6920-7-28

Bendixen, L. D., & Rule, D. C. (2004). An integrative approach to personal epistemology: A guiding model. *Educational Psychologist, 39*(1), 69-80. doi:10.1207/s15326985ep3901_7

Benjamin, L. T., Cavell, T. A., & Shallenberger, W. R. (1984). Staying with initial answers on objective tests: Is it a myth? *Teaching of Psychology, 11*(3), 133-41. doi:10.1177/009862838401100303

Burns, E. R. (2006). Learning syndromes afflicting beginning medical students: Identification and treatment—reflections after forty years of teaching. *Medical Teacher, 28*(3), 230-233. doi:10.1080/01421590600632920

Carey, S. (2000). *Science education as conceptual change* doi://doi-org.ezproxy.library.unlv.edu/10.1016/S0193-3973(99)00046-5

Chang, H., Kang, J., Ham, B., & Lee, Y. (2016). A functional neuroimaging study of the clinical reasoning of medical students. *Advances in Health Sciences Education, 21*(5), 969-982. doi:10.1007/s10459-016-9685-6

Chi, M. T. H. (2005). Commonsense conceptions of emergent processes: Why some misconceptions are robust. *The Journal of the Learning Sciences, 14*(2), 161-199. Retrieved from http://www.jstor.org.ezproxy.library.unlv.edu/stable/25473477

Coderre, S., Wright, B., & McLaughlin, K. (2010). To think is good: Querying an initial hypothesis reduces diagnostic error in medical students. *Academic Medicine : Journal of the Association of American Medical Colleges, 85*(7), 1125-1129. doi:10.1097/ACM.0b013e3181e1b229

Couchman, J. J., Miller, N. E., Zmuda, S. J., Feather, K., & Schwartzmeyer, T. (2016). The instinct fallacy: The metacognition of answering and revising during college exams. *Metacognition and Learning, 11*(2), 171-185. doi:10.1007/s11409-015-9140-8

Darlow, A. L., & Sloman, S. A. (2010). Two systems of reasoning: Architecture and relation to emotion. *Wiley Interdisciplinary Reviews., 1*(3), 382-392. doi:pmid%7E2F26271378

de Neys, W., Rossi, S., & Houde, O. (2013). Bats, balls, and substitution sensitivity: Cognitive misers are no happy fools. *Psychonomic Bulletin and Review, 20*(2), 269-273. doi:10.3758/s13423-013-0384-5

D'Eon, M., Kosmas, C., & Macmillan, J. (2007). Teaching syndromes – A response to learning syndromes: Comments on a paper by Robert Burns (2006) learning syndromes afflicting beginning medical students: Identification and treatment – reflections after forty years of teaching*. *Medical Teacher, 29*(2-3), 280-282. doi:10.1080/01421590701252115

Downing, S. (2005). The effects of violating standard item writing principles on tests and students: The consequences of using flawed test items on achievement examinations in medical education. *Advances in Health Sciences Education, 10*(2), 133-143. doi:10.1007/s10459-004-4019-5

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques. *Psychological Science in the Public Interest, 14*(1), 4-58. doi:10.1177/1529100612453266

Eva, K., W., Link, C., L., Lutfey, K., E., & McKinlay, J., B. (2010). Swapping horses midstream: Factors related to physicians' changing their minds about a diagnosis
. *Academic Medicine, 85*(7), 1112-1117. doi:10.1007/s12467-017-0019-y

Eva, K., W., & Regehr, W., G. (2007). Knowing when to look it up: A new conception of self-assessment ability. *Academic Medicine, 82*(10), S84. doi:10.1097/ACM.0b013e31813e6755

Evans, J. B. T. (2006). The heuristic-analytic theory of reasoning: Extension and evaluation. *Psychonomic Bulletin & Review, 13*(3), 378-395. doi:10.3758/BF03193858

Evans, J. B. T., & Stanovich, K. E. (2013). Dual-process theories of higher cognition. *Perspectives on Psychological Science, 8*(3), 223-241. doi:10.1177/1745691612460685

Ferguson, K. J., Kreiter, C. D., Peterson, M. W., Rowat, J. A., & Elliott, S. T. (2002). Is that your final answer? relationship of changed answers to overall performance on a computer-based medical school course examination. *Teaching and Learning in Medicine, 14*(1), 20-23. doi:10.1207/S15328015TLM1401_6

Fischer, M. R., Herrmann, S., & Kopp, V. (2005). Answering multiple-choice questions in high-stakes medical examinations. *Medical Education, 39*(9), 890-894. doi:10.1111/j.1365-2929.2005.02243.x

Franco, G. M., Muis, K. R., Kendeou, P., Ranellucci, J., Sampasivam, L., & Wang, X. (2012). Examining the influences of epistemic beliefs and knowledge representations on cognitive processing and conceptual change when learning physics. *Learning and Instruction, 22*(1), 62. doi:10.1016/j.learninstruc.2011.06.003

Frederick, S. (2005a). Cognitive reflection and decision making. *Journal of Economic Perspectives, 19*(4), 25-42. doi:10.1257/089533005775196732

Frederick, S. (2005b). Cognitive reflection and decision making. *Journal of Economic Perspectives, 19*(4), 25-42. doi:10.1257/089533005775196732

Geiger, M. A. (1996). On the benefit of changing multiple-choice answers: Student perception and performance. *Education, 117*(1), 108.

Gervais, W. M. (2015). Override the controversy: Analytic thinking predicts endorsement of evolution. *Cognition, 142*, 312.

Gigerenzer, G. (2008). Why heuristics work. *Perspectives on Psychological Science, 3*(1), 20-29. doi:10.1111/j.1745-6916.2008.00058.x

Gigerenzer, G., & Brighton, H. (2009). Homo heuristicus: Why biased minds make better inferences. *Topics in Cognitive Science, 1*(1), 107-143. doi:10.1111/j.1756-8765.2008.01006.x

Handley, S. J., & Trippas, D. (2015). *Dual processes and the interplay between knowledge and structure: A new parallel processing model*. San Diego : doi:10.1016/bs.plm.2014.09.002

Heck, R. H. (2014). In Thomas S. L., Tabata L. N. and Ebooks Corporation (Eds.), *Multilevel and longitudinal modeling with IBM SPSS* (Second edition. ed.) New York : Routledge, Taylor & Francis Group.

Heckler, A. F. (2011). Chapter eight - the role of automatic, bottom-up processes: In the ubiquitous patterns of incorrect answers to science questions. *Psychology of Learning and Motivation, 55*, 227-267. doi://doi-org.ezproxy.library.unlv.edu/10.1016/B978-0-12-387691-1.00008-9

Hess, J., B., Lipner, S., R., Thompson, S., V., Holmboe, L., E., & Graber, L., M. (2015). Blink or think: Can further reflection improve initial diagnostic impressions? *Academic Medicine, 90*(1), 112-118. doi:10.1097/ACM.0000000000000550

Hofer, B. K., & Bendixen, L. D. (2012). Personal epistemology: Theory, research, and future directions. *APA educational psychology handbook, vol 1: Theories, constructs, and critical*

issues (pp. 227-256). US: American Psychological Association. doi:10.1037/13273-009 Retrieved from https://search.proquest.com/docview/1547567514

Hruska, P., Hecker, K., Coderre, S., McLaughlin, K., Cortese, F., Doig, C., . . . Krigolson, O. (2016). Hemispheric activation differences in novice and expert clinicians during clinical decision making. *Advances in Health Sciences Education, 21*(5), 921-933. doi:10.1007/s10459-015-9648-3

Hruska, P., Krigolson, O., Coderre, S., McLaughlin, K., Cortese, F., Doig, C., . . . Hecker, K. (2016). Working memory, reasoning, and expertise in medicine: insights into their relationship using functional neuroimaging. *Advances in Health Sciences Education, 21*(5), 935-952. doi:10.1007/s10459-015-9649-2

Jackson, S. A., Kleitman, S., Howie, P., & Stankov, L. (2016). Cognitive abilities, monitoring confidence, and control thresholds explain individual differences in heuristics and biases. *Frontiers in Psychology, 7*, 1559.

Jeon, M., De Boeck, P., & van der Linden, W. (2017). Modeling answer change behavior: An application of a generalized item response tree model. *Journal of Educational and Behavioral Statistics, 42*(4), 467-490. doi:10.3102/1076998616688015

Kahneman, D. (2011). *Thinking, fast and slow*. New York, NY: Farrar, Straus, and Giroux.

Kahneman, D., & Frederick, S. (2005). A model of heuristic judgment . In K. J. Holyoak, & R. G. Morrison (Eds.), *The cambridge handbook of thinking and reasoning* (pp. 267-293). Cambridge, MA: Cambridge University Press.

Keil, F. C. (2011). Psychology. science starts early. *Science (New York, N.Y.), 331*(6020), 1022. doi:10.1126/science.1195221

Kiat, J. E., Ong, A. R., & Ganesan, A. (2018). The influence of distractor strength and response order on MCQ responding. *Educational Psychology, 38*(3), 368. doi:10.1080/01443410.2017.1349877

Klein, G. A., Calderwood, R., & Clinton-Cirocco, A. (1986). Rapid decision making on the fire ground. *Proceedings of the Human Factors Society Annual Meeting, 30*(6), 576-580. doi:10.1177/154193128603000616

Klien, G. (2015). A naturalistic decision making perspective on studying intuitive decision making. *Journal of Applied Research in Memory and Cognition,* doi:10.1016/j.jarmac.2015.07.001

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments.(statistical data included). *Journal of Personality and Social Psychology, 77*(6), 1121.

Lewandowsky, S., Cook, J., Oberauer, K., Brophy, S., Lloyd, E. A., & Marriott, M. (2015). Recurrent fury: Conspiratorial discourse in the blogosphere triggered by research on the role of conspiracist ideation in climate denial. *Journal of Social and Political Psychology, 3*(1), 142-178. doi:10.5964/jspp.v3i1.443

Lewandowsky, S., Oberauer, K., & Gignac, G. E. (2013). NASA faked the moon Landing—Therefore, (climate) science is a hoax. *Psychological Science, 24*(5), 622-633. doi:10.1177/0956797612457686

Linden, v. d.,Wim J. (2007). A hierarchical framework for modeling speed and accuracy on test items., 287-3123.

Marewski, J. N., Schooler, L. J., & Gigerenzer, G. (2010). *Five principles for studying people's use of heuristics*

Martin, J. W., & Sloman, S. A. (2013). Refining the dual-system theory of choice. *Journal of Consumer Psychology, 23*(4), 552-555. doi:10.1016/j.jcps.2013.04.006

McCain, K. (2015). Explanation and the nature of scientific knowledge. *Science & Education, 24*(7), 827-854. doi:10.1007/s11191-015-9775-5

McConnell, M. M., Regehr, G., Wood, T. J., & Eva, K. W. (2012). Self-monitoring and its relationship to medical knowledge. *Advances in Health Sciences Education, 17*(3), 311-323. doi:10.1007/s10459-011-9305-4

Miller, T. M., & Lindquist, K. (2016). Individual exam analysis using ExamSoft snapshot data. . *Medical Science Educator, 26 (suppl.),* 563.

Monteiro, D., S., Sherbino, D., Jonathan, Ilgen, S., Jonathan, Dore, L., K., Wood, J., T., Young, E., M., . . . Howey, R., E. (2015). Disrupting diagnostic reasoning: Do interruptions, instructions, and experience affect the diagnostic accuracy and response time of residents

and emergency physicians? *Academic Medicine, 90*(4), 511-517. doi:10.1097/ACM.0000000000000614

Moulton, C., Regehr, G., Lingard, L., Merritt, C., & MacRae, H. (2010). "Slowing down when you should": Initiators and influences of the transition from the routine to the effortful. *Journal of Gastrointestinal Surgery, 14*(6), 1019-1026. doi:10.1007/s11605-010-1178-y

Newman, I. R., Gibb, M., & Thompson, V. A. (2017). Rule-based reasoning is fast and belief-based reasoning can be slow: Challenging current explanations of belief-bias and base-rate neglect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 43*(7), 1154-1170. doi:10.1037/xlm0000372

Norman, G., Monteiro, D., S., Sherbino, S., J., Ilgen, G., J., Schmidt, G., H., & Mamede, G., S. (2016). The causes of errors in clinical reasoning: Cognitive biases, knowledge deficits, and dual process thinking. *Academic Medicine, 92*(1) doi:10.1097/ACM.0000000000001421

Norman, G., Sherbino, J., Dore, K., Wood, T., Young, M., Gaissmaier, W., . . . Monteiro, S. (2014). The etiology of diagnostic errors: A controlled trial of system 1 versus system 2 reasoning. *Academic Medicine, 89*(2), 277-284. doi:10.1097/ACM.0000000000000105

Norman, R., G., Monteiro, D., S., Sherbino, S., J., Ilgen, G., J., Schmidt, G., H., & Mamede, G., S. (2017). The causes of errors in clinical reasoning: Cognitive biases, knowledge deficits, and dual process thinking. *Academic Medicine, 92*(1), 23-30. doi:10.1097/ACM.0000000000001421

Pacini, R., & Epstein, S. (1999). The relation of rational and experiential information processing styles to personality, basic beliefs, and the ratio-bias phenomenon. *Journal of Personality and Social Psychology, 76*(6), 972-987. doi:10.1037/0022-3514.76.6.972

Pennycook, G., Cheyne, J. A., Koehler, D. J., & Fugelsang, J. A. (2016). Is the cognitive reflection test a measure of both reflection and intuition? *Behavior Research Methods, 48*(1), 341-348. doi:10.3758/s13428-015-0576-1

Pennycook, G., Ross, R. M., Koehler, D. J., & Fugelsang, J. A. (2017). Dunning-kruger effects in reasoning: Theoretical implications of the failure to recognize incompetence. *Psychonomic Bulletin and Review, 24*(6), 1774-1784. doi:10.3758/s13423-017-1242-7

Pennycook, G., Fugelsang, J. A., & Koehler, D. J. (2015). What makes us think? A three-stage dual-process model of analytic engagement. *Cognitive Psychology, 80*, 34-72. doi:10.1016/j.cogpsych.2015.05.001

Pennycook, G., Trippas, D., Handley, S. J., & Thompson, V. A. (2014). Base rates: Both neglected and intuitive. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*(2), 544-554. doi:10.1037/a0034887

Pretz, J. (2008). Intuition versus analysis: Strategy and experience in complex everyday problem solving. *Memory & Cognition, 36*(3), 554-566. doi:10.3758/MC.36.3.554

Primi, C., Morsanyi, K., Chiesi, F., Donati, M. A., & Hamilton, J. (2016). The development and testing of a new version of the cognitive reflection test applying item response theory (IRT). *Journal of Behavioral Decision Making, 29*(5), 453-469. doi:10.1002/bdm.1883

Rangel, R. H., Möller, L., Sitter, H., Stibane, T., & Strzelczyk, A. (2017). Sure, or unsure? measuring students' confidence and the potential impact on patient safety in multiple-choice questions. *Medical Teacher, 39*(11), 1189-1194. doi:10.1080/0142159X.2017.1362103

Reach, G. (2014). In Ratti C. (Ed.), *Clinical inertia : A critique of medical reason* Cham : Springer.

Rogers, K. (1978). *The gambler*. Los Angeles, CA: United Artists Records.

Rozenblit, L., & Keil, F. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science, 26*(5), 521-562. doi:10.1207/s15516709cog2605_1

Rush, B. R., Rankin, D. C., & White, B. J. (2016). The impact of item-writing flaws and item complexity on examination item difficulty and discrimination value. *BMC Medical Education, 16*(1) doi:10.1186/s12909-016-0773-3

Sherin, B. L., Krakowski, M., & Lee, V. R. (2012). Some assembly required: How scientific explanations are constructed during clinical interviews. *Journal of Research in Science Teaching, 49*(2), 166-198. doi:10.1002/tea.20455

Shtulman, A., & Valcarcel, J. (2012). *Scientific knowledge suppresses but does not supplant earlier intuitions* doi://doi-org.ezproxy.library.unlv.edu/10.1016/j.cognition.2012.04.005

Sinatra, G. M., Kienhues, D., & Hofer, B. K. (2014). Addressing challenges to public understanding of science: Epistemic cognition, motivated reasoning, and conceptual change. *Educational Psychologist, 49*(2), 123-138. doi:10.1080/00461520.2014.916216

Sirota, M., & Juanchich, M. (2018). Effect of response format on cognitive reflection: Validating a two- and four-option multiple choice question version of the cognitive reflection test. *Behavior Research Methods, ,* 1-12. doi:10.3758/s13428-018-1029-4

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin, 119*(1), 3-22. doi:10.1037/0033-2909.119.1.3

Stylianou-Georgiou, A., & Papanastasiou, E. C. (2017). Answer changing in testing situations: The role of metacognition in deciding which answers to review. *Educational Research and Evaluation, ,* 1-17. doi:10.1080/13803611.2017.1390479

Szaszi, B., Szollosi, A., Palfi, B., & Aczel, B. (2017). The cognitive reflection test revisited: Exploring the ways individuals solve the test. *Thinking & Reasoning, 23*(3), 207. doi:10.1080/13546783.2017.1292954

Szollosi, A., Bago, B., Szaszi, B., & Aczel, B. (2017). Exploring the determinants of confidence in the bat-and-ball problem. *Acta Psychologica, 180,* 1-7. doi:10.1016/j.actpsy.2017.08.003

Talukdar, T., Román, F. J., Operskalski, J. T., Zwilling, C. E., & Barbey, A. K. (2018). Individual differences in decision making competence revealed by multivariate fMRI. *Human Brain Mapping, 39*(6), 2664-2672. doi:10.1002/hbm.24032

Toplak, M. E., West, R. F., & Stanovich, K. E. (2014). Assessing miserly information processing: An expansion of the cognitive reflection test. *Thinking and Reasoning, 20*(2), 147-168. doi:10.1080/13546783.2013.844729

Toplak, M. E., West, R. F., & Stanovich, K. E. (2013). Assessing miserly information processing: An expansion of the cognitive reflection test. *Thinking & Reasoning,* , 1-22. doi:10.1080/13546783.2013.844729

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124-1131. doi:10.1126/science.185.4157.1124

van der Linden, W. J. (2009). Conceptual issues in response-time modeling. *Journal of Educational Measurement, 46,* 247-272.

van der Linden, W. J. (2011). Test design and speededness. *Journal of Educational Measurement, 48*(1), 44-60. doi:10.1111/j.1745-3984.2010.00130.x

Wang, T., & Hanson, B. H. (2005). Development and calibration of an item response model that incorporates response time. *Applied Psychological Measurement, 29*(5), 323-339. doi:10.1177/0146621605275984

Wegwarth, O., Gaissmaier, W., & Gigerenzer, G. (2009). Smart strategies for doctors and doctors-in-training: Heuristics in medicine. *Medical Education, 43*(8), 721-728. doi:10.1111/j.1365-2923.2009.03359.x

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education, 18*(2), 163-183. doi:10.1207/s15324818ame1802_2

Zhan, P., Jiao, H., & Liao, D. (2018). Cognitive diagnosis modelling incorporating item response times. *British Journal of Mathematical and Statistical Psychology, 71*(2), 262-286. doi:10.1111/bmsp.12114

Curriculum Vitae

Kristina Lindquist **Email: klindquist1@gmail.com**

Educational Background

2014 (in progress)	Ph.D., University of Nevada, Las Vegas Major: Educational Psychology- Foundations Emphasis: Decision Making and Judgement Dissertation: Beyond Right or Wrong: The influences of thinking disposition and item difficulty on student behavior during high-stakes testing. Anticipated completion: 5/2019
2004 - 2006	M.S., University of Nevada, Las Vegas Major: Kinesiology Emphasis: Motor Behavior and Learning Thesis: A critical review of motor relearning: inferences from neuroscience to application.
1995 - 1999	B.S., University of Massachusetts at Amherst Major: Exercise Science
2/1998-7/1998	Study Abroad, University of Queensland, Brisbane, Australia Focus: Human Movement Studies

Professional Experience

9/17 - present **Director of Instructional Design**

Assistant Professor

Department of Curriculum and Assessment

College of Osteopathic Medicine- Touro University Nevada

- Coordinate with first and second year Assistant Deans of Curriculum and the Senior Associate Dean of Curriculum regarding curriculum schedule
- Develop an exam timeline allowing faculty appropriate time for question submission and revision
- Assemble multi-disciplinary exams for both systems-based and subject-based curricula across the pre-clinical Doctor of Osteopathic Medicine and Master of Medical Health Science programs
- Deliver exams through Examplify (ExamSoft, Inc, 2017) computer-based testing platform
- Interpret individual item results, targeting specific items for additional faculty review
- Provide an overall summary of each assessment to the Senior Associate Dean, Department Chairs, and pertinent course directors
- Evaluate exam and curriculum content in reference to the Comprehensive Osteopathic Medical Licensure EXamination (COMLEX) Blueprint
- Report on assessment of COMLEX Blueprint and American Osteopathic Association Standards for accreditation purposes
- Review student scores and recommend students for meetings with the Academic Success Committee (ASC)
- Meet with struggling students as a member of ASC, recommend strategies to address weaknesses
- Serve as a member of the Continuous Quality Improvement (CQI) committee to identify strengths and weaknesses in the curriculum
- Discuss COMLEX and USMLE board preparation strategies with students, provide individualized counseling for students based on their performance in the curriculum
- Oversee the transition from Blackboard Learning Management System to Canvas Learning Management System for all College of Osteopathic Medicine programs
- Lecture in the first-year Doctor of Osteopathy and Masters of Medical Health Sciences programs

8/16 - 9/17 **Assistant Director**

Office of Academic Services and Institutional Support (OASIS)

Touro University Nevada

- Utilize scientific background to couch information regarding learning strategies within the content that students are experiencing
- Interact with faculty in the College of Medicine and College of Health and Human Services to develop didactic experiences that drive student learning
- Advise students on studying for entry and licensure exams (MCAT, COMLEX Level 1 and 2, USMLE Step 1 and 2, and PANCE)
- Advise students on clinical rotations on studying for shelf/end of rotation exams from the National Board of Osteopathic Medical Examiners (NBOME), National Board of Medical Examiners (NBME), and Physician Assistant Education Association (PAEA)
- Serve on the Academic Success Committee
- Management of OASIS including supervision of the administrative assistant for the office
- Determine (in conjunction with the Director) appropriate and necessary student accommodations for testing and learning in accordance with Federal Americans with Disabilities Act (ADA) guidelines
- Advise students in all programs (College of Medicine and College of Health and Human Services) regarding needs and strategies regarding academic accommodations
- Hire and manage tutors for all courses taught at Touro University Nevada
- Assess effectiveness and report on outcomes of tutoring
- Provide usage reports to the Director of OASIS, Dean of Students, and Vice President of Institutional Effectiveness
- for first and second year students in the Doctor of Osteopathic Medicine program
- Serve as an alternate member of the Student Promotions Committee for students in the Physician Assistant program
- Assess performance of OASIS regarding Institutional Student Learning Objectives (ISLO) for accreditation purposes

8/12 - 9/17 **Learning Specialist**

Office of Academic Services and Institutional Support (OASIS)

Touro University Nevada

- Assist students in developing efficient and effective study strategies based on learning preferences
- Create academic plans for students who are in academic jeopardy
- Develop and deliver student workshops on stress management and study skills
- Liaise with program directors to develop program specific outreach efforts to maximize student use of OASIS resources
- Recruit and hire tutors for the authentic peer tutor and the cohort based tutoring programs
- Assign tutees to tutors and track tutor activity throughout the semester
- Create reports based on number of tutoring hours and grade outcomes for each of the cohorts
- Prepare and apply strategies for Board Certification Tests for students in the Doctor of Osteopathic Medicine, Doctor of Physical Therapy, Master of Occupational Therapy and Master of Physician Assistant programs
- Refer students in distress to the appropriate counseling or other services

10/11-8/12 **Program Director**

Brain Balance Achievement Center

- Assess physical and cognitive skills of children experiencing learning challenges
- Compile reports of assessments for presentation to families of potential students
- Design physical and cognitive programs to address a student's specific learning challenge
- Review program recommendations and assist with at home implementation of success strategies
- Review student progress and report progress to the student and his/her parents
- Advise families regarding nutritional recommendations for student success
- Perform post-program assessments and record changes in physical and/or cognitive skills
- Perform post-program care calls to track student success and advise families regarding student set backs

9/07- 3/11 **Director of Education**

Professional Fitness Institute

- Review, recommend, create, and implement a personal training curriculum to conform with standards set by the National Strength and Conditioning Association (NSCA) and the American Council on Exercise (ACE)
- Develop NSCA-CPT and ACE-CPT exam preparation material
- Recommend strategies for teaching the personal training curriculum at 54 career college campuses nationwide
- Coordinate and schedule presentations, exercise labs, and events for "Personal Training Boot Camp" each month
- Manage a six person team of fitness professionals for the "Personal Training Boot Camp"
- Develop and articulate an educational agenda for Professional Fitness Institute
- Teach lectures and laboratory sessions directed towards producing competent personal trainers
- Advise students regarding test preparation strategies and develop personalized study plans with students

9/06-7/07 **Visiting Instructor**

UNLV Department of Kinesiology, Las Vegas, NV

- Delivery of six upper-level undergraduate Kinesiology classes in the areas of motor behavior, learning, performance enhancement, and human development
- Redesign and update of materials for four courses (KIN 312, 316, 414, 462)
- Review and adopt new textbooks
- Design examinations to appropriately evaluate student progress

Presentations, Publications, and Grants

Academic Presentations:

Miller, T, Lindquist, K (2019). Know when to hold ‘em: Answer changing behaviors in second year medical students. Lecture presentation for the Association for Medical Education in Europe Annual Meeting, August 24-28, 2019 (submitted)

Miller, T., Lindquist, K., Meeks, D., Bloom, A. (2019). Do SOAP note-style question stems offer advantages over vignette-style stems? Lecture presentation for the International Association of Medical Science Educators (IAMSE) Annual Meeting, June 8-11, 2019

Lindquist, K. & Miller, T. (2017). Using individual analysis to improve student test taking skills. Poster presentation for Research Forum hosted by UNLV Graduate and Professional Student Association (GPSA), April 8, 2017.

Miller, T., & Lindquist, K. (2016). Individual exam analysis using Examsoft snapshot data. Lecture presentation for The International Association of Medical Science Educators (IAMSE) Annual Meeting, June 4-7, 2016.

Poliquin, A., Lindquist, K., Miller, T. (2015). Planning for progress: Student success workshop for struggling first and second years. Poster presentation for The American Association of Colleges of Osteopathic Medicine (AACOM) Annual Conference, April 22, 2015.

Lindquist, K., Duford, A., Poliquin, A., Sussman, C. (2015) Is it cheating? Student use of prescription medication for cognitive enhancement. Lecture presentation for National Association of Student Personnel Administrators (NASPA) 2015 National Conference, March 21-25, 2015.

Randall, Y., Poliquin, A., Lindquist, K. (2014) Preparing graduate occupational therapy students for success in anatomy. Poster presentation for The American Occupational Therapy Association (AOTA), April 18, 2015

Randall, Y., Poliquin, A., Lindquist, K., (2014) Preparing graduate students for success in anatomy. Poster presentation for Western Association of Schools and Colleges (WASC) Academic Resources Conference, April 23-25, 2014

Scientific Publications:

Lindquist, K., Guadagnoli, M.A. (2008) Neuroanatomical correlates of motor skill learning: inferences from neuroimaging to behavior. In: M.A. Guadagnoli (ed.) *Human Learning: Biology, Brain, and Neuroscience*

Guadagnoli, M.A., Lindquist, K. (2007) Challenge Point Framework and Efficient Learning for Golf. For: S. Jenkins (ed.) *Annual Review of Golf Coaching*

Guadagnoli, M.A., Grosser L.S., Jones, T.G., Hunt, S., Lindquist, K. (2007) Specificity of feedback and perceived difficulty of a putting task. *Journal of Sport and Exercise Psychology* Vol. 29 (pp. 80-81)

Guadagnoli, M.A., Kohl, R., Lindquist, K., Jones, T.G., Grosser, L.S. (2007) The compatibility of change versus N-change stimulus-response sets. *Journal of Sport and Exercise Psychology* Vol. 29 (pp. 81)

Grants

Miller, T., Lindquist, K. (2018). Lost sleep and lost knowledge? The correlation between sleep and exam performance. Grant submitted to the American Association of Colleges of Osteopathic Medicine (AACOM). Amount funded: $4,970

-Primary duties: Procurement of sleep trackers, recruitment of study participants, informed consent, statistical analysis and interpretation of data

Miller, T., Meeks, D., & Lindquist, K. (2017). A comparison of SOAP note style exam questions to traditional vignette style exam questions. Grant submitted to the International

Association of Medical Science Educators (IAMSE). Amount funded: $4,650

-Primary duties: Statistical analysis and interpretation of item level and student level performance indicators

Lindquist, K., & Miller, T. (2017). Rapids in the river: Downstream effects of challenging questions on student exam performance. Grant submitted to Touro Research Committee. Winner of a 2017 Mentored Student Research Grant Award. Amount funded: $1,500

-Primary duties: Supervision of student worker to perform initial data reduction techniques, statistical analysis and interpretation, presentation and publication of findings

Lindquist, K. (2016). Individual exam analysis using Examsoft snapshot data. Travel grant submitted to the UNLV Graduate and Professional Student Association (GPSA). Amount funded: 1,700

-Travel expenses to the IAMSE Annual Conference in Leiden, Netherlands

Guadagnoli, M. A. (2005). A Common Practices Audit of The First Tee Life Skills Education program. Grant submitted to the First Tee Foundation. Amount funded $92,500.00

-Primary duties: examination of current survey practices of The First Tee, identification of critical information for data collection, proposal of survey methods, presentation of initial investigation findings to The First Tee

Teaching Experience and Lecture Presentations

Yearly Presentations for Touro University Nevada (TUN)

Test Taking Strategies

- Present common misconceptions regarding test questions
- Workshop skills related to test question writing
- Presented to first year Doctor of Osteopathic Medicine students

Planning for Progress: Managing Expectations

- Panel discussion with second year Doctor of Osteopathic Medicine students
- Presented to first year Doctor of Osteopathic Medicine students who had at least 1 fail on exam 1

Relaxation 101

- Introduction to breathing techniques, imagery, and progressive relaxation
- Presented to the TUN student body (open presentation)

Learning Strategies

- Review of scientific support for popular study methods
- Integrated Team Based Learning (TBL) exercise
- Presented to the TUN Medical Health Sciences and Physician Assistant students

Coping with Graduate School Stress

- Overview of the stresses students encounter and strategies students use in graduate school
- Presented to the TUN student body (open presentation)

Planning for Progress: Memory Techniques

- Review of the effects of study strategy on memory consolidation
- Presented to first- and second-year Doctor of Osteopathic Medicine students

Heart Health for Students

- Examination of the American Heart Association's "five tips for heart health" effects on memory and learning

- Presented to the TUN student body (open presentation)

Stress Proof your Holidays!

- Review of the importance of nutrition in supporting learning, memory function, and stress
- Presented to the TUN student body (open presentation)

Do Learners Really Know Best?

- Review of three educational theories: digital natives, learning styles and self-educators
- Presented to the TUN Student Promotions Committee (SPC)

Teaching Experience:

MHSV681: Advanced Study Skills

This course is intended to provide the student with basic and advanced study skills. Instructors will provide evidence-based information regarding study and test taking skills, time management, learning, and communication strategies. Classroom experiences will further prepare students for rigorous courses taken in professional curricula (i.e. Medical and Allied Health Programs).

Course redevelopment (KIN 312, 316, 414, 462):

- Update lecture/course packet content
- Create new exam questions
- Identify appropriate course materials
- Adopt new textbooks
- Implement a final project for each course
- Create grading criteria
- Apply course materials to professional situations

KIN 316 Lifespan Motor Development (Fall '06, Spring '07): Examination of motor and cognitive development throughout the lifespan. Special emphasis on skilled performance, learning theories, motor abilities, individual differences, developmental considerations, and instructional and training procedures for infants through older adulthood. 3 credits.

KIN 414 Enhancing Mental and Motor Abilities (Fall '06): Topics of mental and motor abilities including attention, arousal states, information processing, and practice schedules. Special emphasis on enhancing motor performance through mental strategies. Prerequisites: KIN 250, KIN 312, or KIN 316. 3 credits.

KIN 462 Adult Development and Aging (Spring '07, Fall '08): Physical and psychophysiologic developmental patterns in adulthood and normal aging explored. Relationships of the physical and socio-environmental interactions to the adult physical life process with considerations to successful aging within life stages reviewed. 3 credits.

KIN 312 Motor Control and Learning (Spring '07): Study of the psychomotor domain of movement. Special emphasis on skilled performance, motor learning theories, motor abilities, and motor control. 3 credits.

Lectures presented prior to teaching experience:

3/2006	Arousal states and performance, Kinesiology 414: Enhancement of Mental and Motor Abilities, UNLV
3/2006	Older Adult Health and Fitness, Kinesiology 462: Older Adult Development, UNLV
1/2006	Neuroimaging: A Brief Overview, Kinesiology 462: Older Adult Development, UNLV
1/2006	Cognitive Testing, Kinesiology 462: Older Adult Development, UNLV
11/2005	Older Adult Physiology Overview, Kinesiology 316: Motor Development, UNLV
10/2005	Memory and Learning Models, Kinesiology 316: Motor Development, UNLV

10/2005 Introduction to Motor Behavior, Kinesiology 172: Introduction to Kinesiology, UNLV

7/2005 Introduction to Exercise Physiology, Kinesiology 172: Introduction to Kinesiology, UNLV

5/2005 Older Adult Physiology Overview, Kinesiology 316: Motor Development, UNLV

University Service

2/2018	Champion- Canvas adoption and migration from Blackboard (appointed by Provost)
5/2017	Board Member- Mid Level Professionals (elected position)
2/2017	Touro Strategic Planning Committee- assigned to assess and propose a faculty teaching reward system
11/2014	University Assessment Committee
5/2013	Graduation Speaker Committee

Professional Affiliations

5/2016 International Association of Medical Science Educators (IAMSE)

9/2014 American Educational Research Association (AERA)