Considerations of Reusing Multiple Choice Items to Assess

Medical Certification Repeat Examinees

BY

LISA A. REYES
B.A., University of Chicago, 2006
M.Ed., University of Illinois at Chicago, 2012

DISSERTATION

Submitted as partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Educational Psychology
in the Graduate College of the
University of Illinois at Chicago, 2018

Chicago, Illinois

Defense Committee:

Yue Yin, Chair and Advisor
Everett V. Smith, Jr.
Kimberly A. Lawless
Yoon Soo Park, Medical Education
Tianshu Pan, Pearson

ACKNOWLEDGEMENTS

First, I would like to express my gratitude to Dr. Yue Yin, my advisor and dissertation chair, for her time, support, and expert guidance throughout the dissertation process and my graduate studies overall. I would also like to thank the other members of my dissertation committee, Dr. Everett Smith, Dr. Kimberly Lawless, Dr. Yoon Soo Park, and Dr. Tianshu Pan. Their scholarly guidance greatly contributed to the work presented in this dissertation.

I must express my gratitude to Dr. Craig Deville for his valuable feedback on a preliminary version of the first two chapters of this work and for his continued support of my professional development. Thank you to Dr. Carol M. Myford for her mentorship. Also, I am grateful to the members of my writing group, Kimberly M. Hudson and Tara M. McNaughton. Their writing support helped me get this work off the ground.

Lastly, I would like to acknowledge the friends and family who have supported me throughout the ups and downs of this process. Abigail Crisostomo, Agraja Sharma Guimarães, John Paul Jewell, and Aaron Burton deserve a thank you for their unwavering assurance—and for any snacks they sent to sustain me during long hours at work on this dissertation. I wish to thank my brother Jason, who first taught me how to write with more clarity and confidence. Special thanks to my brother Alex, who introduced me to the field of psychometrics. Finally, my parents contributed to the completion of this dissertation, from their lifelong encouragement of my educational and professional goals to their willingness to dogsit for many hours so I could lock myself away and work on this dissertation. Without them, I would not have made it to the end of this doctoral journey.

LAR

TABLE OF CONTENTS

TABLE OF CONTENTS (continued)

LIST OF TABLES

LIST OF TABLES (continued)

LIST OF FIGURES

LIST OF ABBREVIATIONS

CAT	Computer adaptive testing
CI	Confidence interval
CTT	Classical test theory
SE	Standard error
SEM	Standard error of measurement
SME	Subject matter expert
SP	Standardized patient
TOEFL	Test of English as a Foreign Language

SUMMARY

As gatekeepers to medical specialty practice, medical boards are responsible for accurately identifying who is qualified for certification. Boards often utilize comprehensive multiple choice exams to assess minimum competence and make certification determinations. Generally, boards have retest policies that permit examinees who fail an exam to sit for the exam again. At the same time, boards must frequently reuse exam items due to issues with test equating or limited item availability. Given the potential memory advantages and disadvantages of prior item exposure that might interfere with the assessment of repeat examinees, the ongoing discussion of how best to assess repeat examinees has largely focused on how to minimize item exposure.

The goal of this study was to contribute to the research available to guide the development of defensible retest policies in light of the reuse of multiple choice items. Through investigation of repeat examinee scores on a single medical certification exam, their response patterns on items common across exam attempts, and the measurement capabilities of the items themselves, I aimed to contextualize observed score differences and pass–fail outcomes.

The results indicated that repeat examinee score differences and pass–fail outcomes were largely related to overall content knowledge, rather than any memory effects stemming from prior item exposure. Moreover, indications of the ability to remediate knowledge, the building of false knowledge, and any memory advantages and disadvantages due to prior item exposure varied with category of item difficulty, discrimination power, content area, and cognitive complexity. The findings of this study support the need for medical boards to focus on the quality, not the number, of multiple choice items that they reuse when assessing repeat examinees in order to minimize the potential negative consequences of prior item exposure.

I. INTRODUCTION

A. Background

Medical credentials are requisite designations that patients, health care providers and hospitals, and insurers understand as evidence that a medical practitioner is qualified to safely and competently practice. Medical licensure is required to legally practice medicine. Medical certification in a medical specialty signals that a practitioner has demonstrated sufficient knowledge of that specialty to the appropriate medical board. Medical specialty boards commonly use high-stakes, comprehensive exams to assess minimum competence and make certification pass–fail determinations. Any certification exam must provide valid, reliable measurement of examinee ability to help facilitate accurate pass–fail determinations.

The foremost goal of certification exams is identifying who is sufficiently knowledgeable enough to practice the specialty in order to protect the public (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014; Kane, 1982; Millman, 1989; Smith & Hambleton, 1990). According to classical test theory (CTT), an exam score is based on both the examinee's true proficiency and some amount of error. Such error can interfere with making an accurate certification determination (Haladyna & Downing, 2004; Messick, 1984). The more egregiously inaccurate certification determination is a false positive, meaning that an examinee has passed the exam despite not actually having sufficient knowledge (Clauser & Nungester, 2001; Millman, 1989). Nevertheless, minimizing false negative determinations, failing examinees who indeed have sufficient knowledge for certified practice, is also an important priority for medical boards.

Inaccurate fail determinations can occur due to a number of reasons. Again, all exam scores contain some measurement error and are not perfectly reliable. Minimally competent examinees may struggle with especially difficult questions that happen to be on the exam that day (Millman, 1989). Also, construct-irrelevant factors such as not feeling well or having an inhibiting amount of test anxiety may prevent qualified examinees from demonstrating their true respective level of knowledge.

Even though a board's priority is preventing false positive certifications, inaccurate fail determinations are also problematic for boards. Fail determinations can invite scrutiny of a particular board and its exam development and administration procedures, potentially leading to litigation (Knapp & Knapp, 1995; Mehrens, 1995; Millman, 1989). Also, inaccurate fail determinations simply mean that qualified practitioners are barred from certified practice. Inaccurately barring qualified practitioners can especially hurt some medical specialties by limiting the public's access to practitioners in those specialties (e.g., IHS Inc., 2015; Clauser, Margolis, & Case, 2006; Lupu, 2010).

Board retest policies help address the issue of inaccurate fail determinations by permitting examinees who initially fail a certification exam to sit for the exam again. They provide a pathway to certification for misclassified failing examinees. They also allow examinees who have sufficiently remediated any knowledge gaps since the initial exam attempt to demonstrate their increased knowledge and enter certified practice. Providing an opportunity for medical practitioners to earn an exam score that reflects their true level of knowledge and become board-certified is an important way for boards to uphold their responsibilities to those in their field of practice and to the public.

However, developing appropriate retesting policies presents an array of new issues to ponder and crucial decisions to make. For example, boards must decide if they are going to limit the number of retakes for each examinee. One option might be to limit the number of retakes before enforcing some kind of educational intervention such as an exam review course. Boards may choose to impose a minimum amount of time to pass between exam attempts. Millman (1989) recommended increasing the amount of testing that repeat examinees must pass. If that is not possible, he suggested applying a more rigorous passing standard for those who have previously failed such as averaging the scores on all exam attempts to arrive at a final pass–fail determination or raising the pass point on retakes. Clauser and Nungester (2001) recommended limiting the number of retakes except for in exceptional cases or raising the initial cut score. Clearly, boards have a lot of retesting options to consider.

As gatekeepers to a specific medical specialty, boards are responsible for achieving accurate pass–fail determinations on both initial and repeat exam attempts. Therefore, they must fully understand the implications of any retest procedures they put in place for their respective constituent of examinees. This expectation warrants evaluating the available research regarding repeat examinee performances on both initial and subsequent certification exam attempts as well as the extent to which those performances support their pass–fail outcomes. Without abundant research available to assist boards in performing the necessary due diligence, boards cannot be certain that their retest policies adequately safeguard the integrity of their certification.

B. Statement of the Problem

Developing sound and appropriate retest policies necessitates in-depth research on retest-related issues. In this study, I focused on issues related to reusing multiple choice exam items when retesting examinees. The multiple choice item is the quintessential item format for

measuring different levels of learning in the medical professions (Downing, 2006a, 2006b, 2009b). Also, researchers have frequently discussed the potential challenges with re-administering to repeat examinees multiple choice items they may have seen on their initial exam attempt (e.g., Marsh, Roediger, Bjork, & Bjork, 2007; Raymond, Neustel, & Anderson, 2007; Roediger & Marsh, 2005). Having focused my study on the reuse of multiple choice items to assess the knowledge of repeat examinees, my findings pertain to informing retest policies in the high-stakes medical certification context. Next, I discuss the issues related to multiple choice item development and reuse in medical certification exams.

1. Developing multiple choice items for medical certification exams

A multiple choice item is comprised of two parts: a stem and several response options. The stem poses the question or problem to the examinee. The response options consist of one best answer and several distractors, often two to four (Downing, 2009b; Haladyna, Downing, & Rodriguez, 2002). When presented with a multiple choice item, an examinee must read and process the item, evaluate each of the response options, and apply his or her respective level of content knowledge to select a response. Because an examinee evaluates several response options within only one item, the multiple choice item can be an efficient method of assessing examinee knowledge.

The multiple choice format is ubiquitous in the licensure and certification exam environment because of its efficiency and versatility (Downing, 2006a, 2006b). Multiple choice exams can efficiently and representatively sample a large body of medical knowledge (Clauser et al., 2006; Downing, 2009b). This lends support to the validity of basing pass–fail determinations on exam scores. Therefore, a multiple choice exam that adequately samples all the areas of a medical specialty would be appropriate for identifying who should be certified to practice that

specialty (Kane, 1982, 1994a, 2006). Research intended to inform medical certification retest policies should include a look at the quality of multiple choice items with respect to measuring repeat examinee knowledge across the different content areas and skills of medical practice.

Boards often must invest considerable time and money into training their item writers to write multiple choice items that perform well with regard to measuring the target construct. Abundant guidelines on how to write high quality multiple choice items are available (e.g., Case & Swanson, 2002; Downing, 2006b, 2009b). Haladyna et al. (2002) outlined 31 item-writing guidelines intended to appropriately challenge examinees and facilitate better assessment of their knowledge. Guidelines that are of particular relevance to the current study include that distractors sound plausible and that item writers base them on typical errors they encounter in the field. All the time and expertise required in developing new, high quality multiple choice items for certification exams can amount to approximately $300 to $1,000 per item (Downing, 2006a; Haladyna et al., 2002; Raymond et al., 2007). The bright side of this significant investment is that boards can reuse quality items as long as they remain secure and are not exposed without authorization (Downing, 2006b, 2009b).

2. Reuse of multiple choice items

Due to the high cost of item development, for some boards, reusing items can be essential to the administration and scoring of certification exams. Issues with equating, limited item availability, and exam security may also drive boards to reuse items (Schmeiser & Welch, 2006). For example, boards might have to reuse a considerable proportion of items to facilitate common-item equating, a procedure which provides the same basis of comparison across groups of examinees and thereby allows a board to apply the same pass point across exam administrations. One prevalent guideline states that reused items should comprise a minimum of

20% of the items on an equated exam form (Kolen & Brennan, 2014). Also, boards with limited resources to develop new items might need to reuse items or even repeat entire exam forms. In general, boards might prefer to exercise the option to reuse items that they think test important, relevant medical content.

Unfortunately, the reuse of multiple choice items can interfere with assessing repeat examinees. First, prior item exposure might give repeat examinees an unfair score advantage. If examinees repeat an exam and are given items they had seen during initial exam attempts, they might perform better on these repeated items after having memorized and specifically studied them. In this case, any resulting score gain might be attributable primarily to one's increased familiarity with exam content as opposed to increased knowledge of the field as a whole. This possible score inflation weakens any inference that the examinee is ready to enter certified practice. This risk possibly compounds when a board reuses items across multiple exam forms and permits multiple retakes. In that case, repeat examinee scores will be increasingly based on ability to memorize and study specific items, not on actual professional ability. Therefore, designing retest policies that help minimize item exposure and ensure that any repeat examinees' score gains are not due to prior knowledge of reused items has long been considered crucial to making accurate pass–fail determinations for repeat examinees (e.g., Matthews-Lopez, Woo, Thiemann, Jones, & Gallagher, 2015).

To avert any negative effects of prior item exposure, Rosenfeld, Tannenbaum, and Wesley (1995) suggested policies such as using alternate forms for retesting or using computerized adaptive testing (CAT) to prevent repeat examinees from re-challenging individual items. However, these solutions might not be viable for smaller boards with limited resources to develop enough new items or, particularly in the case of CAT, to meet all the technical

requirements for exam administration. Empirical research on how prior item exposure relates to repeat examinee scores and pass–fail outcomes might then inform more suitable retest policies for such boards.

Another issue with prior item exposure is that it might disproportionately thwart repeat examinees with a rectifiable amount of partial knowledge. Multiple choice items force examinees to evaluate distractors that ideally sound plausible and are based on common errors in the field. The large amount of plausible sounding yet false information appearing on a multiple choice exam might draw some examinees into rationalizing that some of this false information is true (Marsh et al., 2007; Roediger & Butler, 2011; Roediger & Marsh, 2005). For repeat examinees, the reuse of items might translate to repeatedly sampling the very item content that promotes false knowledge and consequently hinder borderline failing examinees who would otherwise be capable of remediating knowledge gaps sufficiently enough for certified practice.

Some might argue that reinforcing false knowledge through the reuse of multiple choice items poses no issue because anyone who succumbs to false knowledge should not be certified anyway. However, multiple choice items can vary with respect to measurement capability. Among items intended to sufficiently challenge certification examinees, some might test important clinical skills, whereas others might inadvertently tap into lower-level thinking about borderline esoteric or trivial medical facts (Clauser et al., 2006; Downing, 2006b; Martinez, 1999; Raymond & Neustel, 2006). I therefore argue that when the reused items that continually stump repeat examinees test less important knowledge or are otherwise flawed, inferences about these repeat examinees' competence are weakened. Therefore, more research is needed regarding whether reusing items contributes to the building of false knowledge among repeat examinees and thereby interferes with the remediation and assessment of their knowledge.

When boards permit retesting, they inherently support the efforts of repeat examinees to demonstrate that they truly qualify for certification. Under this notion, the design of item reuse procedures that neither promote an unfair score advantage, nor misguide otherwise capable repeat examinees from remediating their knowledge gaps, remains an important retesting issue. Empirical research can help enlighten boards as to the consequences of prior item exposure on assessing repeat examinee knowledge. To more thoroughly understand repeat examinee performances on initial and subsequent exam attempts, research needs to address the magnitude of observed score differences, pass rates among repeat examinees, and repeat examinees' performances on unique items versus common items that they have seen on an earlier exam attempt. Research that focuses on these aspects of repeat examinee performance can help boards identify which pitfalls they might face with retesting repeat examinees with reused multiple choice items.

3. Additional issues with the use of multiple choice items to assess repeat examinees

In addition to the effects of prior item exposure on repeat examinee performance, research should more generally address how multiple choice items function among repeat examinees. Prior item exposure as indicated by score gains is the dominant focus of the literature on repeat medical certification testing (e.g., Feinberg, Raymond, & Haist, 2015; O'Neill, Lunz, & Thiede, 2000; Raymond, Neustel, & Anderson, 2007, 2009; O'Neill, Sun, Peabody, & Royal, 2015; Wood, 2009). However, other sources of measurement error can lead to serious certification misclassifications among repeat examinees (Clauser & Nungester, 2001; Millman, 1989). Researching the magnitude of repeat examinees' score gains does not supply all the information that boards need to develop well-guided, defensible retest policies. To build validity for the pass–fail determinations made about repeat examinees, more investigation is

necessary as to how multiple choice items measure knowledge, remediated knowledge, false knowledge, and lack of knowledge among these examinees.

It remains unclear what cognitive processes are at work when repeat examinees respond to multiple choice items. One limitation of the multiple choice item format is that it does not directly reveal the examinee's reasoning process or the degree of confidence that the examinee had in the response he or she selected. For example, exam administrators cannot look at an examinee's multiple choice item responses and conclusively determine whether he or she answered certain items correctly because of guessing or true comprehension. Even when the responses that a repeat examinee supplies are not the result of prior item exposure, boards cannot be sure that such responses accurately reflect the examinee's level of knowledge. Evaluating score gains and pass rates alone will not address this issue. To enhance interpretation of score gains and pass rates among repeat examinees, boards would need to better understand how well their multiple choice items assess repeat examinees. Unfortunately, research is lacking on how medical certification repeat examinees perform on different types of multiple choice items.

Such research is especially important for identifying item flaws that unduly hinder repeat examinees and thus interfere with making accurate pass–fail determinations. Downing (2002, 2006b) found that lower achieving students experienced more difficulty with flawed items compared to higher achieving students. Additionally, flawed items can impact pass–fail rates (Downing, 2006b). Compared to the highest scoring certification examinees, minimally competent repeat examinees might be more prone to struggling with flawed items and consequently earn artificially lower scores. Better understanding of how repeat examinees perform on different types of multiple choice items would help reveal the capabilities and limitations with using certain types of multiple choice items to assess repeat examinees.

The need to better understand repeat examinee performance on different types of items also stems from how items can vary with respect to content, difficulty, and complexity. The particular item content sampled on an exam administration influences scores (Raymond et al., 2007). Therefore, investigating how repeat examinees perform across content areas and item types would strengthen interpretation of their scores as well as the inferences made about the state of their knowledge during initial testing and repeat testing.

Gathering a fuller picture of item performance among repeat examinees is especially crucial for medical boards given the scrutiny and litigation that fail determinations potentially invite (Knapp & Knapp, 1995; Mehrens, 1995). According to Millman (1989), it is reasonable for boards to permit retakes for those who have spent the time and effort to obtain the requisite professional knowledge. This description encompasses every failing examinee. A comprehensive multiple choice exam is only one component of the medical certification process and only one observation of medical competence (Gunderman & Ladowski, 2013). Prior to sitting for the exam, examinees have met firm medical educational requirements. Physicians sitting for certification exams have already become licensed by the National Board of Examiners and the Federation of State Medical Boards to legally practice medicine, and possess supervised real-world experience through residency programs or fellowships (AERA, APA, & NCME, 2014; O'Neill et al., 2015). That many years of medical education and experience would arm failing examinees, especially repeatedly failing examinees, with enough reason to cast suspicion on the exam's capability to measure their accrued knowledge. Further empirical research regarding item performance among repeat examinees can guide retest policies, thereby lending support to the validity of repeat examinee pass–fail determinations and to the defensibility of

retesting procedures. Having research-based retest policies might also help reassure boards as they confront queries, concerns, and complaints from repeat examinees.

Medical boards utilize multiple choice exams to make a binary determination, pass or fail, based on the continuous and complex construct of professional competence. Pass–fail determinations show no distinction between highly or barely competent, nearly or inadequately competent (Clauser & Nungester, 2001). This is why more research focused on repeat examinee performances and item functioning is especially crucial. Such research is needed to serve as the basis for retest policies that facilitate measurement and accurate pass–fail determinations among repeat examinees. To address this need, I examined in this study several different aspects of repeat examinee performance on the certification multiple choice exam of a single medical specialty board.

C. Definitions of Key Terms

In this study, I investigated the assessment of repeat examinee knowledge by examining the relationships between repeat examinee performances on a medical certification multiple choice exam and both prior item exposure and item functioning. An item is *new* when the medical board administered it for the first time, whereas an item is *reused* when the Board had administered it on a previous exam form. A *common item* is a reused item that appeared during both of a repeat examinee's exam attempts. A *unique item* is an item that repeat examinees challenged only once across their respective exam attempts. That is, repeat examinees have had no prior exposure to unique items.

I also use several terms when evaluating response patterns between initial and repeat exam attempts. By *response persistence*, I refer to when an examinee selected the same correct response on both the initial and repeat exam attempts or when an examinee selected the same

distractor on each attempt (i.e., correct–correct or incorrect–same incorrect response patterns). By *response change*, I refer to when an examinee selected a distractor on the initial attempt and the correct answer on the repeat attempt, the correct answer on the initial attempt and a distractor on the repeat attempt, or two different distractors on each attempt (i.e., incorrect–correct, correct–incorrect, incorrect–different incorrect response patterns). By *response time*, I refer to the amount of time that an examinee spent on an individual item, including read the item, selecting a response option, and submitting his or her response. A *change in response time* is the difference between an examinee's response time on an individual item during an initial exam attempt and his or her repeat exam attempt.

D. Purpose of the Study and Research Questions

Boards might design retest policies that better maintain the integrity of their certification and suit their respective examinee populations if they consider how repeat examinees perform during initial testing versus repeat testing. The purpose of this study is to better understand how repeat examinees demonstrate their initial level of knowledge on their first exam attempt and the extent to which they are able to demonstrate sufficiently remediated knowledge on their repeat attempt. To achieve this purpose, I closely investigated repeat examinees' performances on the overall exam, their performances on different types of exam items, and the measurement capabilities of the items themselves. Examining these different aspects of repeat examinee performances facilitated more context around repeat examinees' observed score gains and losses. More specifically, I sought to explain their score differences in terms of ability to remediate knowledge, building of false knowledge, and memory advantages and disadvantages due to prior item exposure. Lastly, I aimed to leverage such insights into repeat examinee performances to

discuss the strength of the inferences made about examinees based on their retest scores. To accomplish these goals, I addressed the following research questions:

1) Which repeat examinees' scores were most amenable to change between initial and repeat exam attempts, and to what extent did score differences indicate sufficiently remediated knowledge?
    a. For repeat examinees who initially borderline failed and examinees who initially clearly failed the exam, do overall exam scores differ significantly between initial and repeat exam attempts?
    b. What is the pass rate among repeat examinees?
    c. Among passing repeat examinees, how many score gains are beyond measurement error?
2) Does examinee performance on different types of items lend support to the pass–fail determinations made about repeat examinees?
    a. How do reused exam items function with respect to distinguishing between different levels of competence among all examinees?
    b. Do repeat examinee subscores, on subtests of items grouped by content area and cognitive complexity level, differ significantly between initial and repeat exam attempt?
3) Do repeat examinee performances on items to which they have had prior exposure indicate any memory advantages or disadvantages from prior item exposure?
    a. Do repeat examinees score differently on common items compared to unique items?

b. How do rates of response persistence and response change compare against changes in response time among all repeat examinees?

c. What are the results from 3(b) specifically among passing repeat examinees with score gains beyond measurement error, as identified in Question 1(c)?

E. Approach of the Study

My research questions were focused on analyzing different aspects of repeat examinee performance on medical certification multiple choice exams. The certification exam data I used in this study are archival data from a psychometric consulting firm that provides psychometric services to outside medical and dental boards. The data come from a single medical specialty board. The data cover a five-year period over which the research participants initially took and then repeated the Board's one-best-answer multiple choice certification exam. I used the dichotomous Rasch model to analyze and score the exam data. Further details on the data collection and data analysis procedures in this study are provided in the Method chapter.

F. Significance of the Study

The available research on medical certification repeat examinee performance has remained limited. Much of the research on medical certification repeat examinees has focused on whether prior item exposure leads to unfair score advantages and the magnitude of observed score gains. The existing research, however, is insufficient for guiding boards in developing appropriate, defensible retest policies because it lacks deeper insights into repeat examinee performances and into the measurement capabilities of the items used to assess examinees' professional competence. The current study has stemmed from my desire to assist medical boards in assuring both the public and their constituents that their assessment processes lead to quality inferences about repeat examinees. Through this study, I offer a closer investigation of

the effects of using and reusing multiple choice items to make pass–fail determinations about repeat examinees.

The existing research, which I more comprehensively discuss in the next chapter, indeed provides valuable implications regarding re-administering items and exam forms to medical certification repeat examinees. My study extends this work by delving more into repeat examinee score differences and item functioning. With this, I have aimed to generate new insights about the reuse of multiple choice items when assessing medical certification repeat examinees.

If boards more deeply understand how repeat examinees perform and how well their items function, they will be more equipped to establish exam procedures that better facilitate measurement of knowledge gaps and remediated knowledge among repeat examinees. Moreover, boards will be better equipped to establish retest policies that hold up to public scrutiny and to litigation. Multiple choice testing often comes under scrutiny across an array of assessment contexts given existing debate regarding how well it can measure examinees' competence in real-world settings (Gunderman & Ladowski, 2013). Professional licensing and certifying organizations face potential litigation when their constituents begin to doubt the quality of their respective exams. Therefore, these organizations need an empirical basis for the retesting policies and procedures they implement. Through this study, I offer findings that medical boards might consider to help minimize inaccurate pass–fail determinations, improve the certification pathway for medical practitioners capable of remediating knowledge gaps, and so uphold the integrity of their certification.

II. REVIEW OF LITERATURE

A. Organization of the Literature Review

This literature review is divided into three main sections, followed by a summary. The first section covers validity and reliability issues that underlie the assessment of repeat examinees' professional competence. Next, I discuss cognition with respect to the assessment of repeat examinees. This section starts with an overview of the cognitive abilities that medical certification multiple choice items might target. This section ends with previous research regarding the effects that taking multiple choice exams can have on recall and knowledge building. Following this overview of cognition-related issues in retesting, I discuss the existing research on repeat examinee performance on high-stakes achievement exams. This section focuses mostly on medical certification repeat examinees.

I relate to each section an existing hypothesis regarding the effects of initial exam exposure on retest scores and outcomes. Figure 1 depicts all of the hypothetical effects I describe. Throughout this chapter, I note the implications of existing theories and findings on future repeat examinee research. The chapter concludes with the collective findings that the existing repeat examinee literature provides as well as the remaining gaps that I aimed to address in the current study.

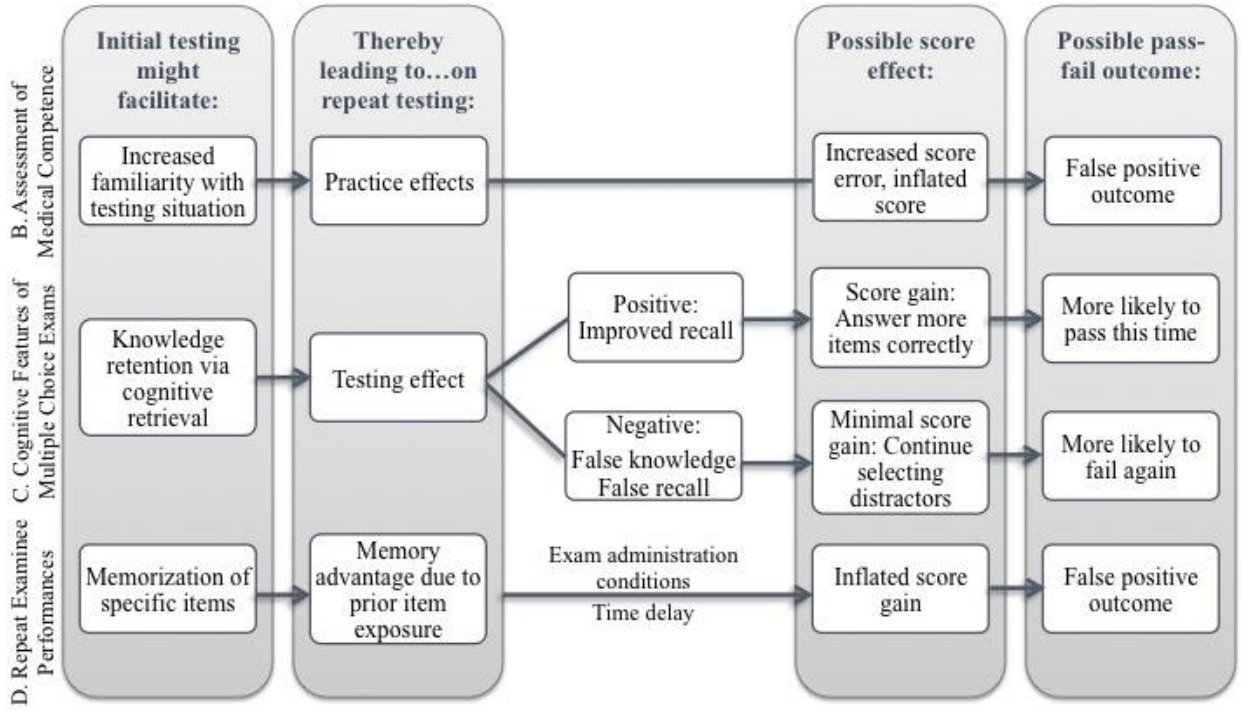

Figure 1. Effects of initial exam exposure on retest scores and outcomes mentioned in the reviewed literature.

B. Assessment of Medical Competence

Licensure and certification boards often regard multiple choice exams as the most valid, reliable, and cost effective exam format for making credentialing determinations (Knapp & Knapp, 1995). Even so, they must take great care in developing exam forms and administration procedures that facilitate valid, reliable measurement. Certain validity and reliability issues may arise when testing repeat examinees. For example, validity issues or random error may artificially raise scores in some cases, suggesting problems with retesting procedures (Raymond & Luciw-Dubas, 2010). In this section, I discuss the literature on the validity and reliability

issues that are particularly relevant to researching repeat examinee behavior and to developing retest policies.

1. Validity of certification determinations

The validity of exam scores used to make medical certification pass–fail determinations is crucial to protecting the public and the profession. Validity relates to the meaningful interpretation of exam scores (e.g., AERA, APA, & NCME, 2014; Cronbach, 1971; Kane, 1994a, 1994b, 2006; Messick, 1984, 1989). In the case of medical certification, validity of the use of exam scores to make certification determinations depends on substantiation for the inferences made regarding examinee medical competence. Within the scope of this study, issues with content validity, construct representation, and construct-irrelevant variance might undermine the inferences made about repeat examinees.

a. Content validity and construct representation

Making inferences regarding future performance in the real-world professional setting is a complex responsibility for medical boards. Boards cannot rely on real-world patient outcomes to assess if someone is qualified for certification because some aspects of the treatment of a patient will always be out of the practitioner's control. As a result, boards should design certification exams to assess readiness to perform the entire range of critical professional responsibilities across different practice settings (Kane, 1982, 1994a; Kane, Kingsbury, Colton, & Estes, 1989; Smith & Hambleton, 1990). That is, boards must design exams that elicit examinees to demonstrate their competence on the full range of important knowledge and skills for practice in order to support the validity of pass–fail determinations based on exam scores (Kane, 1982, 1994a).

Boards can develop multiple choice exams that efficiently assess examinees on a representative sample of the crucial content areas and skills of a medical specialty (Clauser et al., 2006; Downing, 2009b). Certification examinees will fail if they demonstrate that they lack the requisite knowledge at that time. With retesting, failing examinees have at least one more opportunity to demonstrate that they have sufficiently remediated knowledge gaps. However, prior item exposure and other sources of measurement error might unduly influence with their scores. Keeping this in mind, boards might find assurance in the scores and outcomes of examinees who pass on a second or third try if these examinees appear to have demonstrated sufficiently remediated knowledge across all the content areas on the exam. That is, investigating repeat examinee performance differences by content area between initial and repeat exam attempts can provide clarity on overall exam score differences and pass–fail outcomes among repeat examinees.

The validity of scores and pass–fail determinations for repeat examinees also depends on the extent to which the exam prompts repeat examinees to demonstrate that they can critically apply their knowledge across a wide array of patient conditions and medical settings. That is, a certification exam should include items that have been written to tap into higher cognitive abilities, such as critical thinking, and that have been shown to distinguish between different levels of competence. This is why evaluating repeat examinee performance on different types of items remains important to interpreting repeat examinee scores and pass–fail outcomes.

b. Construct-irrelevant variance

Repeat examinees might be particularly vulnerable to issues with construct-irrelevant variance that undermines the validity of scores and outcomes. One potential source of construct-irrelevance is item quality. High quality items are attained through effective item

writing. Such items minimize construct-irrelevant variance, thereby reducing measurement error and increasing exam validity. On the other hand, poorly written, flawed items result in construct-irrelevant variance that impedes measurement and the meaningful interpretation of exam scores (Downing, 2002). Flawed items also have the potential to negatively impact pass–fail rates (Downing, 2002, 2006b).

In one study, Downing (2002) found that flawed items artificially made exams more difficult. Flawed items were more difficult for the lower achieving students compared to the higher achieving students in the study. Moreover, flawed items resulted in higher fail rates, particularly when the pass point was at or just above the center of the exam score distribution. Therefore, item quality might disproportionately thwart failing examinees, particularly borderline failing examinees. This disadvantage brought to borderline failing examinees highlights the need to investigate the quality of the items used to assess repeat examinees, specifically how well these items function for measuring examinees of different levels of ability.

Another potential source of construct-irrelevant variance concerns the actual sample of items on the exam form and prior item exposure. Reusing items when assessing repeat examinees can result in an exam of items biased toward items that repeat examinees have seen before (O'Neill et al., 2000). Score gains on a repeat attempt should be due to a remediated knowledge base, not due to improved performance only on the previously seen items. Therefore, reusing items might threaten the validity of scores and pass–fail determinations on a repeat exam attempt. Accordingly, comparing repeat examinee scores on common items versus unique items might shed light on the implications of reusing multiple choice items to assess repeat examinees.

2. Reliability and measurement error

Repeat examinee certification testing gives rise to issues of reliability and measurement error. Reduced reliability and increased measurement error can reduce classification accuracy (Clauser et al., 2006). With a medical certification exam, classification accuracy is essential. All certification exam scores are associated with some amount of measurement error. Classification accuracy means that the pass–fail classifications based on those exam scores are nevertheless in agreement with examinees' true levels of professional competence.

Because a certification pass–fail classification is based on demonstrating minimum competence, certification exams require items that yield the most measurement precision near the pass point. Examinees whose true ability levels are near the pass point are more likely to be misclassified because of the errors of measurement associated with exam scores near the pass point (Hambleton & Slater, 1997; Kane, 1996). A borderline examinee who barely passes or fails on one exam attempt will have about a 50-50 probability of either passing or failing on another exam attempt (Downing, 2006b). Therefore, errors that impact examinees closest to the pass point are more likely to lead to inaccurate pass–fail determinations. Researching classification accuracy with respect to assessing repeat examinees therefore requires special attention to both the score differences and the item functioning, especially among borderline examinees.

Classification consistency, the reliability of the pass–fail classifications made about examinees across parallel exam administrations, is also a priority in the certification context (Hambleton & Slater, 1997). With classification consistency, passing examinees should pass again if retesting within a short time frame, and failing examinees should fail again. A

certification exam must produce consistent pass–fail classifications in order to justify the use of exam scores to assess professional competence (Kane, 1996). The costs of poor medical certification classifications are high. Poor classifications can ultimately hurt public safety, public access to qualified medical practitioners, the career trajectories of the affected examinees, and the professional reputation of the medical board involved.

If a certification exam produces consistent classifications, only certain types of examinees would be expected to pass a repeat exam attempt. The first type of repeat examinee is one whose initial exam score fell just below the pass point due to measurement error. In this case, the examinee was initially misclassified, and retesting helped resolve the misclassification. The second type of repeat examinee is one who has indeed remediated knowledge gaps sufficiently enough to meet minimum competence. In this case, it would seem more realistic to expect that examinees whose initial scores were within striking distance of the pass point, as opposed to well below the pass point, are more likely to sufficiently remediate their knowledge and qualify for certified practice.

Many have expressed concern about false positive classifications among repeat examinees. For example, some have theorized that allowing multiple exam retakes increases probability that repeat examinees will receive false positive certifications (Clauser et al., 2006; Clauser & Nungester, 2001; Millman, 1989). That is, each time a near competent examinee attempts the exam, the possibility that he or she passes due to measurement error increases. Given concerns about false positive certifications among repeat examinees, evaluating the strength of the pass–fail determinations made about repeat examinees should entail close investigation of who passes repeat exam attempts and who does so with score gains beyond measurement error.

Score reliability is also crucial to the quality of a certification exam. Score reliability refers to the reproducibility of exam scores across administrations, and lack of reliability can undermine validity (Haladyna & Downing, 2004). The reliability of exam scores serves as an indicator of the classification accuracy and classification consistency of a certification exam. Because measurement errors can affect score interpretation and reduce reliability, minimizing errors is important to ensuring that exam items distinguish between different levels of examinee competence to help facilitate accurate, reproducible pass–fail outcomes.

Random measurement error alone will not cause someone whose initial score is well below the pass point to eventually pass a certification exam. Examinees with scores well below the pass point might be expected to score higher upon retesting as their scores regress toward the mean exam score (Rosenfeld et al., 1995). However, regardless of number of exam attempts, examinees with scores well below the pass point will likely continue to fail the exam (Millman, 1989). In fact, some certification examinees never pass their exams (Gershon, 2005). Observing dramatic score differences between exam attempts might signal that more serious sources of score error, such as prior item exposure, are afoot.

That is, some measurement errors have the potential to seriously impact the inferences made about repeat examinee proficiency. In the next section, I describe several sources of error that frequently appear in the literature on repeat examinees. These errors might lead to undue score differences among repeat examinees.

a. Content sampling

Regarding reliability, content sampling can be a source of measurement error because the specific sample of items appearing on an exam form can influence examinee scores (Raymond & Luciw-Dubas, 2010; Raymond et al., 2007). On an initial exam attempt, an

examinee might be challenged with a personally difficult set of items (Millman, 1989). With the reuse of items, repeat examinees might then be repeatedly challenged with the same personally difficult items, perhaps constituting a greater obstacle to certification than if they receive a fresh sample of items (Feinberg et al., 2015). Looking more closely at repeat examinee score differences and response patterns might indicate to what an extent a particular sample of items on an exam form lends to observed scores and outcomes.

b. Item quality

Not only does item quality impact validity, but it also impacts measurement error and thus reliability. Specifically, poorly written, flawed items can introduce error that lowers the reliability of pass–fail determinations (Downing, 2009b). Ensuring reliability means administering high quality items that minimize the chance that examinees who indeed have the knowledge answer an item incorrectly, or that examinees who do not possess the knowledge answer the items correctly (Kane, 1982). An analysis of item performance can help determine whether certification exam items are of high quality, are written clearly enough for examinees, and are contributing to the reliability of the exam.

c. Guessing

Guessing is a random measurement error of frequent concern in high-stakes testing environments. Examinees are more likely to guess than to skip an item when they are not penalized for incorrect answers (Hutchinson, 1982). When examinees are permitted to guess, there is a difference between how many items an examinee must answer correctly to pass the exam versus how many items the examinee must know the answer to, with the former number being notably lower (Millman, 1989). Therefore, lucky guessing can lead to inflated scores, thereby undermining the validity of scores and any inferences based on those scores.

However, concerns about blind guessing might be overstated in the literature. Challenging, thoroughly reviewed and edited items can hinder lucky guessing (Downing, 2009b). Even if an examinee makes the occasional lucky guess, an examinee is unlikely to guess his or her way to a passing score on a long, well-written exam. Additionally, guessing generally involves partial knowledge. That is, when examinees are not certain of the correct answer to an item, they likely to engage in informed guessing, applying their partial knowledge to eliminate response options and deduce the correct answer. Informed guessing entails partial knowledge and uncertainty of the correct answer, whereas blind guessing entails complete lack of knowledge and inability to recognize the correct answer at all (Hutchinson, 1982). Downing (2009b) asserted that health practitioners often must rely on partial knowledge in real practice and that exams can still yield information about examinee ability even when the examinee uses partial knowledge to make an informed guess. An informed guess that leads to the correct answer is less of a threat to reliability.

On the other hand, guessing might have one particular pitfall for repeat examinees. Roediger and Marsh (2005) found that examinees will guess when they first come upon unfamiliar item content. On subsequent exam retakes, these examinees will often persist in their incorrect guesses, perhaps because they had already established a rationale for those guesses. Therefore, guessing might increase faulty reasoning and thereby increase the potential for the building of false knowledge over repeated exam attempts. This false knowledge building exemplifies one type of memory disadvantage due to prior item exposure.

Repeat examinees might engage in different guessing behaviors, and the effect of different guessing behaviors on repeat examinee scores might vary. Therefore, analyzing repeat examinee response patterns during initial and repeat testing might indicate which guessing

behaviors repeat examinees are exhibiting and help explain whether the detrimental or less egregious effects of guessing relate to observed scores.

d. Practice effects

Practice effects are exam score gains simply due an examinee's previous exposure to an exam rather than to actual increases in knowledge. For example, initial exposure might result in a reduction in test anxiety, improved test-taking skills, or greater familiarity with item format, increased confidence, or a combination of these. These consequences would then influence subsequent exam performance. For repeat examinees, practice effects might introduce measurement error, compromise exam scores, and thereby increase false positive and false negative pass–fail decisions (Rosenfeld et al., 1995).

Kulik, Kulik, and Bangert (1984) conducted a meta-analytic study from 40 studies on how taking a practice test impacts subsequent test performance. The majority of the studies focused on aptitude tests, though several were about achievement tests. The results indicated that students generally performed better when they first took practice tests. The size of the effect was greater when the subsequent exam was an identical form compared to a parallel form or when students took a greater number of practice tests beforehand. Also, the size of the effect was greater for students of high ability compared to those of low ability. The researchers theorized that higher ability students may have been better able to learn from the practice test, identify their weaknesses, and improve their knowledge than lower ability students.

There are some key differences to keep in mind when attempting to apply these findings to high-stakes certification retesting. First, practice tests are lower-stakes. Some students in the studies may have used the practice tests as a starting point or pretest prior to instruction. With that, especially high gains on subsequent testing are expected. With high-stakes exams,

examinees would assumedly come as prepared as possible. Also, most of the tests included in the study were aptitude tests. It might be reasonable to expect gains just due to students' mental development over time. Moreover, compared to the highest scoring certification examinees, nearly competent examinees might have a more difficult time identifying and remediating knowledge gaps. Therefore, practice effects might manifest differently in the high-stakes certification context.

This suggests a continued need to determine whether practice effects are strong enough to propel certification repeat examinee scores above the pass point. Determining this might well warrant comparison of repeat examinee performances on different types of items. The impact of practice effects does not appear similar across all item types (Rosenfeld et al., 1995; Wing, 1980). Therefore, seeing how different item characteristics lend to reduced practice effects in certification testing might help boards determine how best to mitigate negative consequences from repeat testing.

C. Cognitive Features of Multiple Choice Exams

Exams prompt examinees to engage in different cognitive processes to respond to items. In this part of the chapter, I discuss studies on the cognitive processes involved in test-taking and the consequences on retention and knowledge building among repeat examinees.

1. Cognitive abilities measured by multiple choice items

To facilitate accurate pass–fail determinations, medical certification exams should reflect the range of cognitive demands in the field (Kane, 1982; Martinez, 1999). This means presenting items that prompt examinees to engage in the relevant thinking abilities in order to supply answers. Higher cognitive items, items that tap into higher order thinking skills, would

appropriately challenge certification examinees to demonstrate whether their respective ability to carry out medical procedures is sufficient.

Despite, or due to, the widespread use of multiple choice items in testing, the multiple choice item format has been widely critiqued. One criticism of multiple choice items stems from research suggesting that, because these items present the correct answer to examinees, they tap only into recognition processes and are less effective than constructed response items at tapping into the recall processes that foster long-term retention (Carpenter & Delosh, 2006; Foos & Fisher, 1988; Glover, 1989; Ozuru, Briner, Kurby, & McNamara, 2013). Other research has indicated that well-written multiple choice items that require examinees to evaluate plausible distractors might indeed prompt recall processes that support long-term retention of correct answers and related information, at least as effectively as the constructed response format (Little & Bjork, 2015; Little, Bjork, Bjork, & Angello, 2012). Though multiple choice items may routinely tap into recognition instead of recall, in this paper, I use the term "recall" rather than "recognition" to refer to multiple choice exam items intended to engage cognitive retrieval of isolated facts. This is in order to be consistent with the terminology that the medical specialty board in this study used when they originally developed the certification exam items.

Another criticism of multiple choice items has been that they neglect to test higher levels of thinking (e.g., Cronbach, 1988). However, researchers have found that multiple choice items can be written to challenge cognitive abilities at higher levels (Downing, 2006b, 2009b; Hancock, 1994; Martinez, 1999). For example, a multiple choice item might present a vignette about a clinical patient and then prompt the examinee to select which response option would be the best next step in managing the treatment of the patient. This type of item is a higher cognitive item. It requires examinees not only to recall factual information but also to correctly

interpret the information provided in the stem and apply their knowledge to select a response (Anderson & Krathwohl, 2001; Bloom, Englehart, Furst, Hill, & Krathwohl, 1956). An exam will ideally contain an appropriate mix of higher cognitive items and items that require only recall (Downing, 2009b). That way, boards are able to determine which examinees can meet the full range of cognitive demands in medical practice.

Even though multiple choice items can assess higher cognitive ability and complex proficiency, in practice they too often target lower level cognitive abilities (Martinez, 1999). This might be due to how difficult it is to effectively write higher cognitive items. Even when item writers set out to do so, writing cognitively challenging clinical items is difficult (Martinez, 1999; Raymond & Neustel, 2006). Item writers must exercise tremendous skill in writing multiple choice stems and distractors that prompt higher levels of thinking such as interpretation and application (Hancock, 1994). On the other hand, writing lower cognitive items such as recall items remains easier. Consequently, item writers might instead produce lower cognitive items to assess medical certification examinees, which can inadvertently result in testing trivial or esoteric content knowledge (Downing, 2006b; Martinez, 1999; Raymond & Neustel, 2006). This can lead to lower exam scores and increased measurement error (Downing, 2002).

Multiple choice items can test important clinical problem-solving and application skills, and they can test trivial medical information (Downing, 2006b; Martinez, 1999; Raymond & Neustel, 2006). Knowing the correct answers to cognitively challenging items and not just to recall items can add meaning to one's scores. If passing repeat examinees are found to demonstrate proficiency with cognitively more complex items, boards can be better assured that repeat examinees' score gains are due to improved competence as opposed to measurement error. That is, looking at the extent to which repeat examinees demonstrate their abilities on both lower

and higher cognitive items during exam attempts can help reinforce pass–fail determinations based on scores.

2. Memorial effects from taking multiple choice exams

Several researchers have suggested that the information presented in multiple choice items can take root in examinees' memory and become part of their respective knowledge base. The cognitive process of retrieving past learning from memory to supply answers on an initial test often improves mastery and recall, sometimes more so than directly studying the content (Carpenter, 2009). That is, testing not only measures knowledge but can also shape it. This is referred to in the literature as the testing effect (e.g., Fazio, Agarwal, Marsh & Roediger, 2010; Glover, 1989; Marsh et al., 2007; Roediger & Butler, 2011; Roediger & Marsh, 2005).

The consequences associated with the testing effect can be either beneficial or detrimental to examinee knowledge. Through retesting, repeated retrieval and the testing effect can contribute to long-term retention of content (Fazio et al., 2010; Roediger & Butler, 2011). However, multiple choice items present distractors featuring plausible sounding false information to examinees. Even if examinees ultimately select the correct answer during initial testing, repetition of a plausible distractor might cause some examinees to eventually falsely recall the distractor as being true (Hasher, Goldstein, & Toppino, 1977). When examinees are not already certain of the correct answer, they might apply somewhat faulty reasoning to mistakenly deduce that one of the plausible distractors is correct. This process of rationalizing a distractor can lead examinees to fold the distractor's false information into their existing knowledge bases and reproduce this false knowledge on subsequent exam retakes (Marsh et al., 2007; Roediger & Butler, 2011; Roediger & Marsh, 2005).

Roediger and Marsh (2005) found some indication of this false knowledge building when comparing performances on a reading comprehension multiple choice initial and final test taken on the same day, with a brief filler task between the tests. In their study, they had half of the research participants read one set of 18 passages and the other half read a second set of 18 passages. Both the initial and final tests given to all participants contained items on the complete set of 36 passages. On the initial test, examinees were told they had to guess and select a response even if they were unsure of the correct answer. On the final test, they were strongly discouraged from guessing and told to draw a line in the response space if they were unsure of the correct answer.

Their results showed some evidence of the testing effect, with examinees generally answering more items correctly on the final test compared to the initial test. That is, initial testing helped prepare examinees for the final test. However, 75% of all incorrect responses that examinees selected on the final test were the same distractors they selected on the initial test. The different instructions on guessing between the initial and final tests would suggest that a larger proportion of the responses on the initial test were selected with some degree of uncertainly compared to those on the final test. One potential explanation is that some examinees may have selected distractors that they believed to be correct and so did not bother reviewing after the initial test. Another possibility is that some examinees who unsuccessfully deliberated on items during the initial test may have gone on to regard the distractors they had selected as true information. They then supplied these distractors as answers on the final test even when instructed not to guess. Either way, examinees often reselected distractors on the final test instead of sufficiently addressing their knowledge gaps to ensure they arrived at the correct answers on the final test.

Marsh et al. (2007) observed a similar phenomenon. As an initial test, they had college undergraduate students respond to four subject-specific tests containing SAT II items in biology, chemistry, U.S. history, and world history. They used standard SAT II instructions on these initial tests, which penalizes incorrect answers and thereby discourages blind guessing. The students did not receive any performance feedback on these initial tests. Next, the students took a final general knowledge test containing both new items and items testing the same concepts as the initial SAT II tests. For the final test, the researchers revised some of the SAT II recall items to become application items, or items in which the students had to apply their knowledge of some concept in order to respond. Again, researchers observed a large positive testing effect. The students were more likely to answer final test questions correctly when the content of those questions appeared on the initial tests. However, researchers also noted a negative testing effect. The students were also more likely to select the same distractors on the final test as they did on the initial test, suggesting they neither realized nor addressed their initial knowledge gaps. The researchers also found that switching a recall item to an application item reduced the likelihood that students would select a distractor, but it did not eliminate the negative testing effect altogether. The researchers explained that even though prior exposure to the SAT II format and items was generally beneficial, it might have also led to examinees incorporating distractors into their respective knowledge bases. The study therefore suggests that faulty reasoning on multiple choice items that tap into higher level thinking can lead to false knowledge building, though the benefits of initial testing typically outweigh this negative consequence.

Repeat certification examinees might be particularly susceptible to the negative memorial effects of multiple choice testing. Armed with only partial or no knowledge of item content, they might read plausible sounding yet incorrect distractors during their initial exam attempt. When

they select such distractors, the process of rationalizing those distractors might lead to the accumulation of false knowledge. Even when initially answering correctly, some distractors might sound increasingly reasonable to them and lead to false recall of the answers at a later time. By the time examinees retake the certification exam, they might use false knowledge to reselect distractors or even succumb to false recall and switch from a correct to an incorrect response option. For multiple repeat examinees, false knowledge and false recall originating from prior exam exposure might be greater through the increased confrontation with misinformation. For these examinees, challenging an identical exam form or at least an exam form containing a high proportion of reused items might mean multiple exposures to the same misinformation, thereby firmly planting false knowledge and false recall.

Butler and Roediger (2008) have found that this multiple choice testing pitfall can be mitigated by the use of item-level feedback on examinee performance. They stated that such feedback should at least include supplying examinees with the correct responses so that examinees might persist in their correct reasoning and rectify faulty reasoning. However, this level of feedback is often not feasible in the medical certification context (Gunderman & Ladowski, 2013). As boards sometimes need to reuse items, supplying examinees with the correct answers would compromise the security of a board's item bank and might undermine scores for subsequent examinees. Boards are then often relegated to providing more general feedback on exam performance, by content area for example. Whereas this approach might provide examinees with some idea of their strengths and weaknesses across the larger knowledge domain of their medical specialty, it would not directly identify which specific gaps in knowledge each examinee needs to remedy. In fact, some test takers might misuse or misinterpret content-level feedback by, for example, over-interpreting the results or trying to

piece together content scores to determine how close they came to the pass point even though total test scores cannot be calculated from individual content area subscores (McCallin, 2016). Some examinees might even believe they have mastered a particular concept when they actually have not (Gunderman & Ladowski, 2013; Roediger & Marsh, 2005). Therefore, repeat examinees with a lot of specific content gaps have limited guidance on what errors to focus their remediation efforts so they might retake the exam successfully.

Generally, taking an exam can serve as a form of learning, outweighing the potential pitfall of false knowledge building (Marsh et al., 2007; Roediger & Marsh, 2005). However, it might have a different result on repeat examinees. For repeat examinees, repeatedly rationalizing plausible sounding distractors might promote misinformation and thereby false knowledge building (Roediger & Marsh, 2005). Though Millman (1989) and Clauser and Nungester (2001) asserted that repeat exam attempts increase the risk for false positive classifications, the body of research on the memorial effects of multiple choice testing suggests that repeated item exposure might thwart some near competent examinees in their efforts to remediate their knowledge. If the process of taking an exam indeed alters knowledge, then examining how this might manifest in the retest performances of repeat examinees remains important.

D. Repeat Examinee Performances on High-Stakes Achievement Tests

This part of the chapter is a review of existing empirical research on repeat examinee performance on high-stakes achievement exams, with a focus on medical certification exams. Though most of the included studies offer some consensus regarding the effects of prior item exposure, they point to several different directions for future research on repeat examinee behaviors and performances.

1. Non-medical exams

Wilson (1987) studied score differences among repeat examinees on the Test of English as a Foreign Language (TOEFL) exam, a high-stakes exam that universities use for admissions for international applicants. The TOEFL exam contains multiple choice items as well as a speaking section and a writing section. Wilson studied two samples of examinees to compare long-term versus short-term changes in test performance. The first sample consisted of examinees with English proficiency at all levels who retested within two to five years after the initial test. The second sample consisted of examinees enrolled in formal instruction in English as a second or foreign language after their initial TOEFL scores placed them there. They retested within one to 12 months after the initial test. Wilson did not indicate the proportion of items reused across exam attempts. The results showed that examinee scores substantially increased with each retest. Also, longer time delays positively predicted considerable score gains among the repeat examinees. This suggests that longer time delays gave repeat examinees more time to improve their English language facility.

In contrast are the results from the Geving, Webb, and Davis (2005) study on examinees who repeated a real estate licensure exam. Over 9,000 repeat examinees were included in this study. The participants were permitted to retake the exam as often as they wanted within a six-month period, and the number of retakes for participants ranged from one to nine. No exam form was repeated, though on average about 12% of the items on each exam were reused items. Examinees received feedback on how many items they answered correctly per content area, though they did not receive any feedback on individual items.

The results indicated that there was no correlation between number of retakes and score change. Generally, pass rates among repeaters fell after two retakes. Those who had to take the

exam more than two times had difficulty increasing their knowledge enough to eventually pass, and examinees who needed the fewest retakes experienced the largest score gains. Examinees whose scores were extremely low on earlier attempts generally improved their scores with each attempt, but often this was not enough to pass. These results counter concerns that the number of retakes can increase the probability of passing due to measurement error (Clauser et al., 2006; Clauser & Nungester, 2001; Millman, 1989).

The researchers also found that seeing the same items on multiple exam attempts did not increase scores. On average, 64% of examinees responded correctly to an item the first time, but only 59% correctly responded to that same item the second time. Examinees responded with the same distractors 26.57% of the time, changed to an incorrect option 14.14% of the time, and changed to the correct answer 9% of the time. Therefore, these examinees did not experience any unfair score advantages due to prior item exposure.

The average time delay was 25 days between attempts. Similar to what Wilson (1987) observed with TOEFL repeat examinees, as number of days between attempts increased, score gain increased by a small but statistically significant amount. A longer time delay may have meant that examinees were taking the time to remediate their knowledge gaps and apply new information. This would lend support to medical certification retest policies that impose a minimum time interval between exam attempts.

2. Medical credentialing exams

The existing studies on medical certification repeat examinees have focused on several different aspects of repeat examinee performance. As a result, they have yielded different types of findings. For clarity, this part of the chapter is divided by type of finding, or theme.

a. Score gains and score advantages

The dominant focus in the literature on medical certification repeat examinees has been on whether examinees benefit from prior item exposure and experience unfair score advantages. Prior item exposure might pose a major problem to medical boards given that score gains due to prior item exposure would invalidate exam results. However, most of the existing research collectively suggests that even though repeat examinee scores often increase on subsequent testing, repeat examinees generally do not benefit from any unfair score advantages.

First, O'Neill et al. (2000) looked at repeat examinees on a medical technology CAT exam and found that an examinee's prior item exposure resulted in a small, statistically significant increase on a CAT exam. However, this benefit was small enough to fall well within the bounds of measurement error. For some examinees, re-challenging items led to decreased scores overall. The researchers noted the CAT format's item selection algorithm and item pool depth might have protected against the negative effects of prior item exposure.

As for a non-adaptive certification exam, Raymond et al. (2007) investigated the effects of administering identical and parallel exam forms to repeat examinees for two different radiologic technologic certification programs. Samples of repeat examinees were randomly assigned either the identical form or the parallel form. For both certification exams, repeat examinees generally experienced substantial score gains over their initial scores. However, score gains by examinees who retested with an identical form were indistinguishable from score gains by examinees who retested with a parallel form. These results suggest that boards might not need to worry about unfair score advantages among repeat examinees if they administer the same exam form more than once.

In a 2009 follow-up study, Raymond et al. again randomly assigned radiologic technologic repeat examinees to either an identical or parallel exam forms during retesting. They observed that mean score gains were comparable between both groups of repeat examinees. Therefore, prior exposure to the entire exam form did not yield any major score advantage. This study lends additional support to the conclusion that boards need not be too concerned that prior item exposure will result in unfair score advantages among repeat examinees.

Before indiscriminately applying these findings, boards should investigate score gains by repeat examinees who test under their specific testing procedures. In both studies, the initial and repeat exam attempts for both exams were completed within a calendar year, and the researchers looked at only one year of exams for both exams. A longer term study based on more exam administrations might also reveal more useful information about how to detect and minimize the effects of prior exposure on identical and parallel form repeat testing, particularly for certification programs that test only annually.

Wood (2009) similarly explored the effects of prior exposure on repeat exam performance for a Canadian screening examination for medical physicians about to enter supervised medical practice. Instead of researching the re-administration of identical forms, he limited the number of reused items to 36 items and randomly mixed them with 288 new items. The results showed that repeat examinees achieved similar score increases on the reused and new items alike. Repeat examinees generally benefitted from having taken the exam before, but prior item exposure did not result in an unfair score advantage for them.

More recently, Feinberg et al. (2015) found additional support for earlier findings that prior exposure generally did not unfairly advantage medical certification repeat examinees. For this sample of 388 repeat examinees, there was a one-year delay between attempts. They found

that initial and repeat exam scores were positively correlated (r = .63). However, on either identical or parallel forms, they found no relationship between scores on initial attempt and gain scores on repeat attempt, indicating that initial performance did not moderate score improvement. They also found no discernible differences in score gains between repeat examinees who received an identical form compared to those who received a parallel form containing 25% reused items from the prior year.

Contrary to the findings of the aforementioned research studies, O'Neill et al. (2015) found that prior item exposure may have inflated some scores. In their study, 988 examinees took the American Board of Family Medicine's certification exam twice within a single year, failing the initial attempt and retaking the exam five months later. The examinees were randomly assigned one of two exam forms for their initial attempt and then challenged the other form on their repeat attempt. Each exam form consisted of 260 items, 99 of which appeared on both forms. Performing repeated measures *t* tests on the initial and repeat exam scores, the researchers found that the mean score increase from the initial attempt to the repeat attempt was positive and significantly significant. On the initial exam attempt, the mean difference between examinee scores on common and new items was positive but not statistically significant. On the second attempt, however, the mean difference between common and new items was not only positive but also statistically significant. Therefore, score gains may have been due to a general increase in content knowledge, though a small but noticeable increase might be due to prior item exposure. This finding differs from the preceding study that I discussed, illustrating the need to continually investigate how prior item exposure might impact different populations of repeat examinees.

In this same study, the researchers also found that approximately 3% of repeat examinees passed with higher scores on common items compared to on unique items (O'Neill et al., 2015). This proportion represents examinees whose passing scores may have been inflated due to prior item exposure. However, the researchers mentioned neither how many of these examinees earned scores just under the pass point on their initial attempts nor the actual pass–fail rate among all repeat examinees. Without this information, it remains difficult to more fully understand the impact of memory advantages and disadvantages from prior item exposure on repeat examinees' classifications.

Collectively, the previous studies on score gains among repeat examinees on multiple choice medical certification exams show general consensus that prior item exposure has a limited effect on scores. The majority of repeat examinees did not benefit enough to pass. Researchers have observed similar repeat examinee behaviors in the certification performance assessment setting. On a certification performance assessment like an oral exam, the clinical cases that examinees are presented with might be more memorable. Rather than responding to a long, comprehensive multiple choice exam, performance assessment examinees act out professional tasks and work on the same content for a longer period of time. They are also presented with fewer cases total, and thus their exposure to each case is fairly lengthy. Therefore, it might be reasonable to expect that repeat performance assessment examinees have an unfair advantage when they see the same clinical content on a subsequent exam. However, researchers have found little indication that prior exposure resulted in inflated scores on a performance assessment retake (e.g., Boulet, McKinley, Whelan, & Hambleton, 2003; Swygert, Balog, & Jobe, 2010).

For example, Boulet et al. (2003) investigated repeat examinee performance on a licensure exam that uses standardized patients (SPs) who act out clinical vignettes to help assess

physicians' clinical knowledge and skills. During the exam, examinees encountered a total of 10 SPs for 15 minutes each. They acted out the tasks they would typically carry out with actual patients such as gathering data, performing physical exams, and taking patient notes. On average, repeat examinees generally obtained higher scores during retesting. However, encountering the same SP or clinical vignette during retesting was not associated with score gains. That is, repeat exposure to specific exam elements did not notably impact scores.

The results were similar when Swygert et al. (2010) looked at repeat examinees for the United States Medical Licensing Examination series Step 2 Clinical Skills exam, a high-stakes performance assessment that uses SPs. The researchers observed significant mean score gains on the second attempt. However, they did not find any significant gains for examinees who encountered repeat SPs or scenarios. In fact, for one area on the exam, examinees who encountered repeat exam elements actually performed worse on average. The researchers noted this might be because some examinees were confused about encountering the same content or were overconfident because they were already familiar with the content. Another possibility is that examinees may have experienced increased anxiety if they recognized that they were being re-challenged with content that was difficult for them on the initial attempt. This increased anxiety may have negatively impacted scores. Either way, these results suggest that prior item exposure does not necessarily inflate scores. In fact, prior exposure can sometimes be detrimental.

The existing literature collectively indicates that even though medical certification repeat examinees generally performed better on repeat exam attempts, most repeat examinees did not benefit from any unfair score advantage due to seeing the same items during retesting. However, not all of the studies I mentioned in this section were in unanimous agreement about the impact

of prior item exposure. Consequently, there remains the need to regularly investigate repeat examinee performances as well as the need to study different medical certification repeat testing populations. Additionally, the researchers of these studies generally looked at all repeat examinees in their respective samples as a whole. Dividing the repeat examinees by how close they were to the pass point on their initial attempt would have helped contextualize observed score gains. It remains unclear whether, for example, the examinees who experienced the greatest score gains had borderline failing versus clearly failing scores on their initial attempt. Comparing score differences by initial performance might shed more light on repeat examinee score differences and on the quality of exam items used to assess their professional competence.

b. Pass rates

In addition to analyzing score gains among repeat examinees, it is crucial to determine how often observed score gains were sufficient enough for repeat examinees to pass. After all, the pass–fail determination is the ultimate product of high-stakes medical licensure and certification exams. Generally, researchers have found that pass rates among repeat examinees are lower than that among first-time examinees. For example, O'Neill et al. (2000) found that about 42% of repeat examinees passed. Those examinees earned scores near the pass point when they first took the exam. It is important to note that the CAT format, involved in this study, has a less restricted supply of items ready to target the pass point. A CAT exam can continuously administer such items to an examinee if his or her ability estimate keeps hovering around the pass point, until reaching a pass–fail decision with a predetermined level of confidence. In contrast, repeat examinees might be more likely to see the same items on a non-adaptive exam.

As for repeat examinee pass rates on non-adaptive exams, pass rates among repeat examinees were generally considerably lower than pass rates among first-time examinees

(Feinberg et al., 2015; Raymond & Luciw-Dubas 2010; Raymond et al., 2007). Wood (2009) found that the average initial exam score for passing repeat examinees was higher than that for failing repeat examinees. Raymond et al. (2007) found that pass rate among repeat examinees who took an identical form of one of the exams they analyzed (52.0%) was generally slightly higher compared to that among those who took a parallel form (49.0%), though this difference was not statistically significantly different. (On the other exam that they investigated, they similarly found a higher pass rate among those who took the identical form (68.4%) compared to those who took the parallel form (51.2%). However, the mean initial attempt score of those who received an identical form was somewhat higher than that for those receiving a parallel form, so the difference in pass rates could have stemmed from who comprised each form group.) In contrast, Feinberg et al. (2015) observed a higher pass rate among examinees who retested on a parallel exam form (61.0%) compared to those who retested on an identical form (51.2%).

The results of these studies suggest that a large proportion of examinees have the same pass–fail outcome on a repeat exam attempt. Even though repeat examinees generally scored higher during retesting, they generally did not score high enough to pass. Because some of the researchers neglected to indicate the initial scores of the repeat examinees who passed, it remains difficult to interpret these pass rates. For example, it is challenging to determine if anyone who passed was nearly proficient on the initial attempt or to find indication that no individual passed due to prior item exposure. Moreover, the aforementioned studies diverge on whether more passing repeat examinees retested on an identical exam form or on a parallel form. Knowing how many repeat examinees pass is insufficient to guiding retest policies. Information on which repeat examinees pass and which fail is also needed to better evaluate the pass–fail

determinations among repeat examinees and to gauge the extent to which certain repeat examinees are able to address knowledge gaps.

c. Multiple repeat exam attempts

Wood (2009) compared the performance of examinees who retested once versus examinees who retested multiple times. The mean scores for single and multiple repeat examinees on the reused items were not statistically different. This suggests that both single and multiple repeat examinees experienced roughly equal score gains on reused items. Therefore, repeated exposure to the same items did not compound any effects of prior item exposure, either by giving repeat examinees an unfair score advantage or by prompting them to build false knowledge.

These findings, as well as the findings of the Geving et al. (2005) study on real estate licensure examinees, help dispute concerns that allowing multiple exam retakes can increase the risk of false positive certifications (Clauser & Nungester, 2001; Millman, 1989). Unfortunately, medical certification studies that address the performances of multiple repeat examinees on multiple choice items are scarce. Additional research is needed regarding multiple repeat testing and the longer term effects of prior item exposure, though this might be difficult as many medical specialty boards might not have a sufficient number of multiple-take repeaters to study.

d. Time delays between exam attempts

When attempting to understand the effects of prior item exposure, researchers have considered the relationship between time interval between exam attempts and exam score differences. A common concern with retesting has been that prior item exposure might allow examinees to memorize and specifically study reused exam content, leading to inflated, undue

score gains. The time between exposure and retesting might then mediate how well repeat examinees are able to memorize exam content and use it to their advantage.

Previous medical certification research has not entirely addressed the role of time interval between repeat exam attempts. Researchers outside the medical certification context have found that the testing effect can persist over a delay between retakes (Fazio et al., 2010; Roediger & Butler, 2011). However, Raymond et al. (2007, 2009) found that a time interval as short as three weeks between exam attempts had no significant effect on medical certification repeat examinees' score gains. Because the focus of the existing research on medical certification retesting has been prior item exposure alone, the literature has not sufficiently addressed whether longer time delays lead to improved scores on medical certification exams as such delays did for some non-certification exams (Geving et al., 2015; Wilson, 1987).

e. Differences in exam response patterns

Investigating response patterns on initial versus repeat exam attempts can clarify score differences by aiding comparison of repeat examinees' initial versus later levels of knowledge. Given the personal and professional consequences that failing a certification exam might have on individual examinees, it would be reasonable to assume that repeat examinees spend a lot of time and effort to improve their knowledge. However, the research on repeat examinee response patterns suggests that repeat examinees generally persisted in selecting the same distractors when responding to common items.

When comparing the response patterns on reused questions between initial and repeat exam attempts, Wood (2009) found that repeat examinees tended to choose the same response option on repeat attempts, including choosing the same incorrect option both times. Incorrect–correct response changes (17% of responses) occurred less frequently than persistently incorrect

responses (21% of responses). Also, 13% of responses were correct–incorrect response changes. These findings indicate that repeat examinees did not overwhelmingly benefit from prior item exposure. They did not resolve prior mistakes.

Feinberg et al. (2015) compared examinee responses between identical forms and similarly found that repeat examinees frequently selected the same incorrect response option on the second attempt. When responses were incorrect on both attempts, repeat examinees selected the same incorrect option 68% of the time. Also, 12% of repeat examinee responses were correct–incorrect changes. In this case, repeat examinees were able to remediate their knowledge on only some of the common items.

O'Neill et al. (2015) also compared the response patterns on common items. On average, about 15% of the responses were incorrect–correct changes. Like the previous researchers, they found that repeat examinees tended to choose the same response option on both exam attempts. This includes selecting the same incorrect response about 18% of the time. Furthermore, about 7% of responses were incorrect–different incorrect changes, and approximately 11% were correct–incorrect changes. Once again, repeat examinees were able to improve their performance on some common items, but they did not remedy all of their knowledge gaps.

Altogether these studies indicate that repeat examinees persisted in their incorrect answers. However, the thought processes that guided their responses on common items are still unclear. For example, correct–incorrect changes could have been due to guessing on initial and repeat exam attempts (O'Neill et al., 2015). Conversely, applying the findings from researchers like Marsh et al. (2007), repeat examinees may have sometimes incorporated distractor information into their knowledge bases. It is also unclear if, for example, some of the items on an exam were flawed in a way that prompted repeat examinees to choose the same distractors.

Perhaps the common items that repeat examinees answered incorrectly were also markedly difficult for passing examinees. More research on item quality would help clarify this. Also, future research should examine the possibility that repeatedly sampling the same content and repeatedly presenting the same misinformation might hinder repeat examinees' remediation attempts.

f. Differences in response times

Analyzing differences in item response times against multiple choice response patterns can bolster understanding of the cognitive processes of examinees during retesting. Response times might proxy the confidence that repeat examinees had in their responses, blind guessing, or informed deliberation (Lasry, Watkins, Mazur, & Ibrahim, 2013). Raymond et al. (2009) compared total testing time between first and second exam attempts and between identical and parallel forms. They found that examinees who retested on the identical form had significantly shorter total response times than those who retested on the parallel form. However, score gains were comparable between both groups of repeat examinees. That is, even if examinees recognized and more quickly responded to items they saw on their initial exam attempt, their prior exposure to those items did not necessarily produce any unfair score advantage.

Feinberg et al. (2015) examined changes in response times by individual item. They found that repeat examinees generally took less time to respond to an item correctly than incorrectly, regardless of whether they were answering reused or new items. This suggests the repeat examinees may have generally felt more confident in their correct responses on common and new items alike, perhaps due to actual gains in content knowledge. However, prior exposure

may have resulted in some advantage as the greatest decrease in item response time occurred with incorrect–correct response changes.

With computer-delivered exams, comparing response times across individual items during initial and repeat attempts might provide more insight regarding repeat examinees' respective level of knowledge and the cognitive processes in which they were engaging when responding to multiple choice items. On common items, researchers might look closely at the relationships between performing better or worse on an item, response persistence or change, and changes in response time to help discern specific test-taking behaviors such as guessing or recognizing the answer to a previously seen item. For example, improved performance and a considerably shorter response time on a common item might suggest a memory advantage from prior item exposure or remediated knowledge. An incorrect–different incorrect response change and a shorter response time might suggest inability to remediate knowledge, guessing, or both. Lastly, a correct–incorrect response change and a comparable response time might suggest false recall due to the repetition of plausible sounding misinformation.

E. Summary of Related Literature

The previous research studies regarding repeat examinee performances on medical certification multiple choice exams provide some important insight into the consequences of reusing items to test repeat examinees. The researchers generally agreed that any advantages of prior item exposure were relatively modest. This pattern might be unsurprising as examinees who did not demonstrate sufficient knowledge and thus failed their initial exam attempt might be less capable of capitalizing from prior item exposure on their repeat attempt (Kulik et al., 1984; Feinberg et al., 2015). If they lacked sufficient knowledge the first time, remediation attempts might not adequately help because they might again lack it the second time. Some of the

researchers included in this chapter advised that because identical-form repeat examinees generally did not experience any inflated score gains due to prior item exposure, boards might spare themselves the labor, time, and expense producing a parallel form for repeat examinees if their resources are limited.

However, medical specialty boards should not interpret this suggestion as a blanket recommendation. Specific testing procedures can also differentially impact score gains and advantages for repeat examinees. For example, Raymond et al. (2007) remarked that certain testing conditions present in their study might explain why any advantages of taking an identical form were minimal for the repeat examinees. First, the exams contained many items administered in random sequence, which could have hindered attempts to memorize items and response options and to immediately recognize them when retesting. Second, repeat examinees had to wait at least three weeks before exam attempts, which may have further thwarted memorization efforts. Lastly, examinees received performance feedback only by general content area. Therefore, many examinees may not have been aware of the content knowledge gaps they needed to remedy in order to pass a repeat attempt. Altogether, these possible explanations of minimal advantages for the repeat examinees show that certain testing procedures might help mitigate the advantages of prior item exposure and practice effects. When designing retest policies regarding the reuse of multiple choice items, boards need to take into account how their specific testing and score reporting procedures might affect their respective population of repeat examinees.

Furthermore, additional research is needed to better contextualize repeat examinee score gains and pass–fail rates and thereby explain the strength of repeat examinees' score inferences. For example, comparing how often borderline failing versus clearly failing examinees pass their

repeat attempts would indicate whether measurement errors might be interfering with pass–fail outcomes. On the other hand, verifying that most passing repeat examinees were nearly competent during their initial attempts would suggest that examinees who had passed truly improved their knowledge, not just benefitted from prior item exposure. Therefore, comparing pass–fail changes between repeat examinees who borderline failed and examinees who clearly failed their initial attempts will help shed light on the extent to which repeat examinees remediated knowledge gaps as opposed to benefitted from measurement errors.

Examining item performance via examinee response patterns would also help further contextualize score differences. Analyzing rates of response change along with changes in item response times can help gauge the extent to which repeat examinees are able to identify and remediate their respective knowledge gaps, experience the memory effects of prior item exposure, or both. Moreover, analyzing how reused items function across all first-time certification examinees, including non-repeaters, might help clarify the measurement capabilities of the multiple choice items used to repeatedly assess repeat examinees. Because repeat examinees might be thwarted by issues such as content sampling and false knowledge building, it remains important to investigate the quality of the items administered to them more than once.

Investigating performance on different types of items might also shed light on the strength of the exam score inferences for repeat medical certification examinees. Confirming that exam items are functioning well can reassure boards of the items' capability to assess repeat examinee competence. In addition, investigating how repeat examinees perform across different content areas and across items of differing cognitive complexity can help explain their observed scores and lend support to their pass–fail outcomes. Deciphering when a repeat examinee luckily guessed or truly improved his or her knowledge is difficult when only looking at his or

her selected responses. Taking a closer look at performance on different types of items can help confirm whether passing repeat examinees are demonstrating sufficiently remediated knowledge and improved critical thinking skills.

The goal of this study was to address the need to better contextualize repeat examinees' score differences between initial and repeat testing and their pass–fail retest outcomes by exploring different aspects of their exam performances. This study serves as another case study of repeat examinee score differences and pass–fail rates, thereby extending the literature on medical certification repeat examinees. In this study, I compared changes in score in light of initial exam performance to identify which repeat examinees showed remediable knowledge gaps and to investigate how measurement error related to initial and retest pass–fail outcomes. I also analyzed repeat examinees' performances and response patterns on different types of items on the certification exam. By helping to uncover how well the exam items measured the full range of content knowledge and cognitive demands of the medical specialty, results from this part of my study might well aid interpretation of score differences and pass–fail rates among repeat examinees. Studying these aspects of repeat examinee performance has allowed me to help bridge the existing research on the memorial effects of testing with the existing empirical research on score gains among medical certification repeat examinees. With that, I have aimed to build upon the available research that medical boards can use to help develop well-guided, defensible retest policies that improve the pathway to certification for repeat examinees while safeguarding the integrity of their certification.

As discussed in the Introduction chapter, I addressed the following research questions during this study:

1) Which repeat examinees' scores were most amenable to change between initial and repeat exam attempts, and to what extent did score differences indicate sufficiently remediated knowledge?
    a. For repeat examinees who initially borderline failed and examinees who initially clearly failed the exam, do overall exam scores differ significantly between initial and repeat exam attempts?
    b. What is the pass rate among repeat examinees?
    c. Among passing repeat examinees, how many score gains are beyond measurement error?
2) Does examinee performance on different types of items lend support to the pass–fail determinations made about repeat examinees?
    a. How do reused exam items function with respect to distinguishing between different levels of competence among all examinees?
    b. Do repeat examinee subscores, on subtests of items grouped by content area and cognitive complexity level, differ significantly between initial and repeat exam attempt?
3) Do repeat examinee performances on items to which they have had prior exposure indicate any memory advantages or disadvantages from prior item exposure?
    a. Do repeat examinees score differently on common items compared to unique items?
    b. How do rates of response persistence and response change compare against changes in response time among all repeat examinees?

c. What are the results from 3(b) specifically among passing repeat examinees with score gains beyond measurement error, as identified in Question 1(c)?

III. METHOD

In this chapter, I describe the data and the methods that I used in this study. The purpose of my study was to better understand how repeat examinees demonstrated their initial level of knowledge on their first exam attempt and the extent to which they were able to demonstrate sufficiently remediated knowledge on their repeat attempt. To contextualize their score differences between their initial and repeat exam attempts, I analyzed their performances on the overall exam, their performances on different types of exam items, and the measurement capabilities of the items themselves. By examining these different aspects of their exam performances, I aimed to study their observed score differences and pass–fail outcomes, in essence evaluating the reuse of multiple choice items to make pass–fail determinations about repeat examinees.

A. Data Source

The certification exam data I used in this study are archival data from a psychometric consulting firm that serves medical and dental boards. The data come from a single medical specialty board in the United States. The data consist of certification exam items, responses, and item response times from a six-year period, the year of the benchmark exam plus the five-year period of interest in this study. The Office for the Protection of Research Subjects at the University of Illinois at Chicago determined that this research does not involve human subjects and therefore does not require Institutional Review Board (IRB) approval or exemption.

To keep details of this board's certification exam confidential, I have not disclosed any specific details regarding the content of the Board's exam forms, nor do I provide personal information about the certification examinees or members of the Board.

B. Participants

The dataset in this study contains the responses and response times, in seconds, of 924 individual examinees during the five-year period of interest. However, only 62 of those examinees (6.7%) initially took and repeated the exam within that period. These 62 repeat examinees comprised the sample of repeat examinees in this study, though I analyzed the responses from all examinees in the dataset to obtain more stable estimates of examinee ability and item difficulty. Some of the examinees in the repeat sample retook the exam multiple times over the five-year period. Of the 62 repeat examinees, 45 examinees (72.6%) retook the exam only once. Eleven examinees (17.7%) repeated the exam twice, four (6.5%) repeated the exam three times, and two (3.2%) repeated the exam four times. For this study, I focused on their initial exam attempts and their first repeat attempts only.

A sizeable proportion of the examinees in the dataset, 60 examinees, had initially taken the exam prior to the five-year period and then took the exam more than once during the five-year period. Though they took the exam multiple times over the five-year period, I excluded them from the repeat examinee sample in this study. Due to the limitations of the archival data, I could not determine when these examinees initially took the exam, if they repeated the exam at least once prior to the five-year period, and to which individual items they had prior exposure. I retained their responses when obtaining overall exam scores and item difficulty values but did not use their scores to study repeat examinee performance. The exclusion of their scores considerably reduced the sample size of repeat examinees in my study. Nevertheless, this compromise was needed to have available information on the response patterns, response times, and item content for all repeat examinees in this study.

To be eligible to sit for the certification exam, examinees were required to have a Doctor of Medicine (MD) degree and a current medical license. They were required to have successfully completed a residency, followed by the completion of an accredited training fellowship program in the Board's medical specialty. Therefore, the certification examinees were rather homogenous in terms of medical education and clinical practice backgrounds. No demographic information about the examinees is available as the Board did not collect this information.

C. Exam Development

Subject matter experts (SMEs) in the specialty were responsible for the development of the one-best-answer multiple choice certification exam, creating the content outline and serving as item writers. The Board selected a representative committee of SMEs based on their professional history, reputation, and expertise. All SMEs were longtime Board-certified practitioners, and many were fellowship program directors. No demographic information about the SMEs is available.

When creating the content outline, SMEs ensured that each exam form would representatively sample all aspects of clinical practice in the specialty. They first divided the exam content into two major areas, basic biomedical science and clinical science. Basic biomedical science encompasses basic science knowledge such as foundational biological concepts and scientific principles that are pertinent to practice within the medical specialty. Clinical science encompasses the knowledge of medicine needed to practice the specialty in the clinical setting. After dividing the content into these two major content areas, board SMEs then divided those areas into a total of over 10 topic categories specific to their medical specialty.

They then determined how to distribute items across each of these categories so as to accurately reflect the critical capabilities of safe, effective practice in the medical specialty.

Prior to writing items, the SMEs underwent training on how to properly write multiple choice items. They were trained to use multiple choice item writing guidelines such as those I mentioned in the Introduction chapter. They were also encouraged to write fewer recall items and more cognitively complex items, such as interpretation items or clinical problem-solving items. Nevertheless, they still needed to write recall items to test recognition of certain science concepts and cover all content topics. For each exam form, the SMEs independently wrote new items and then met as a group to revise those items. Lastly, as a group, they selected new and previously administered items to fill content guidelines and put on each exam form.

Using the benchmark exam, the exam that immediately preceded the five-year period in my study, a panel of SMEs selected by the Board established the benchmark scale and the exam pass point. They used a variant of the Angoff (1971/1984) method to establish a criterion-referenced pass point. Under this method, SMEs independently estimated the proportion of minimally competent examinees, those who should just pass the exam and be certified, who would answer each item on the benchmark exam correctly. The sum of the averages over all exam items over all SMEs served as the raw passing score, or the expected score for minimally competent examinees. The standard setting panel then adjusted the raw passing score upward to create a more stringent pass point and reduce the likelihood that no examinee will pass due to measurement error. In light of Millman's (1989) concern that any downward adjustment of a passing score will dilute the passing standard and increase the likelihood of false positive certifications, I specifically chose to study this certification exam because the overseeing board

had decided to increase the raw passing score and thereby create a more rigorous passing standard.

The exams I analyzed in this study were linked to this benchmark scale through common-item equating, which allows comparisons between the different exam administrations over the five-year period (Kolen & Brennan, 2014). For the first exam year in this study, the exam form contained 350 items. The four subsequent exam forms contained 300 items. All five exam forms had at least 20% overlap or at least 30 common items to facilitate common-item equating to the benchmark scale (Kolen & Brennan, 2014).

D. Data Collection

The medical board administered its multiple choice certification exam once a year over the five-year period. The exam was delivered via computer at secure, proctored testing centers throughout the United States. Examinees were given the items in random sequence. They had a time limit of seven hours to complete the exam.

Because the exam was given annually, the minimum interval of time between initial and repeat exam attempts was one year. The longest time interval among research participants was three years. Of the 62 repeat examinees in this study, 53 examinees (85.5%) had a time interval of one year, seven examinees (11.3%) had an interval of two years, and two examinees (3.2%) had an interval of three years.

After each exam administration and before exam scoring, a committee of SMEs from the Board conducted an item analysis using CTT statistics as indicators of item quality. Specifically, they reviewed item p-values and point-biserial correlations. The p-value is the proportion of examinees who responded correctly to the item, and it is an indication item difficulty. Too low of a p-value suggests that the item might be too challenging to assess minimum competence.

The point-biserial correlation for an item, symbolized by r_{pbis}, is the correlation between the selection of a particular response option for that item and examinee raw scores on the overall exam. A point-biserial correlation for the correct response indicates how well the item discriminates between more and less able examinees. The committee evaluated exam items with low p-values, low point-biserial correlations for the correct responses, or distractors with a point-biserial correlation at least as strong as that of the correct response. They then determined which items to remove from scoring. Across the five-year period in this study, this item review process resulted in a range of about 3% to 11% of administered items being removed from scoring on each exam. I removed these same items from my study as SMEs had already deemed them faulty.

After exam scoring, each examinee received a score report. The score report included each examinee's pass–fail result and total exam score. For informational purposes, the report also listed the examinee's subscores by exam content area and topic category. The score report did not provide any item-level performance feedback.

E. Data Analyses

In this section, I describe the methods of data analysis that I carried out to answer my research questions. I have divided this section by main research question.

1. Research Question 1

Which repeat examinees' scores were most amenable to change between initial and repeat exam attempts, and to what extent did score differences indicate sufficiently remediated knowledge?

a. For repeat examinees who initially borderline failed and examinees who initially clearly failed the exam, do overall exam scores differ significantly between initial and repeat exam attempts?
b. What is the pass rate among repeat examinees?
c. Among passing repeat examinees, how many score gains are beyond measurement error?

To answer research Question 1, I first applied the dichotomous Rasch (1960) model to derive examinee scores on the overall exam as this model was used when the certification examinees in the archival dataset were originally scored. I used the software program Winsteps 3.69 to analyze the exam response data (Linacre, 2009b). The Rasch model represents the interaction between examinees and items. It models the probability that an examinee (*n*) will correctly answer an item (*i*) as a function of the difference between examinee ability (θ_n) and item difficulty (δ_i) (Wright & Stone, 1979). This model is expressed by the equation (1)

$$P_{ni} = \frac{exp(\theta_n - \delta_i)}{1+exp(\theta_n - \delta_i)} \quad (1)$$

where

P_{ni} = probability of examinee *n* answering item *i* correctly,

θ_n = level of ability of examinee *n*, and

δ_i = difficulty of item *i*.

Provided that the data adequately fit the Rasch model, Rasch analysis reports examinee ability estimates and item difficulty estimates on the same interval scale in the same units and allows direct comparisons between ability and difficulty measures (Bond & Fox, 2007). The scale score units are called log-odd units or logits.

For this part of the study, I conducted a Rasch analysis for each exam administration to derive examinee scores, using common-item equating to link the five exams to the same benchmark scale. Putting all scores on the same scale facilitated direct comparison of different examinees from different exam administrations without having to consider which specific exam forms they took (Angoff, 1971/1984; Kolen & Brennan, 2014). That is, a repeat examinee's score was not dependent on when he or she took the exam during the five-year period. Using equated scaled scores also allowed me to apply the Board's pass point to all five exams.

Each exam form had a proportion of reused items with difficulty values already calibrated to the benchmark scale. Under common-item equating, those difficulty calibrations served as anchor values. When scoring a given administration, I linked the administration to the benchmark scale by fixing the difficulties of the reused items to the anchor values. While carrying out the common-item equating, I ensured that the anchor items were sufficiently consistent in their benchmark-calibrated difficulty measures across administrations. That is, I unanchored any items that did not maintain their difficulty on a subsequent exam. For each anchored item, the Winsteps software provides a displacement statistic that estimates the difference in difficulty between the anchor value and what the difficulty estimate would be if the item had not been anchored (Linacre, 2009a). To evaluate these displacement statistics, I used the absolute value greater than or equal to 1.0 logit as the displacement threshold for unanchoring common items. The use of 1.0 logits as a displacement statistic is in keeping with the operational scoring procedures when the certification examinees in the archival dataset were originally scored.

After obtaining the overall exam scores using Rasch analysis and common-item equating, I answered Question 1(a) by two types of analysis, one to evaluate how the sample of repeat

examinees performed as a whole and another to evaluate how they performed individually. For the group-level analysis, I conducted a repeated measures *t* test between initial attempt and repeat attempt scores to investigate whether repeat examinees as a whole experienced overall score increases during the repeat exam attempt. Because this was the first of seven group-level comparisons I conducted on the same set of exam response data for this study, I used the Holm-Bonferroni method to adjust the family-wise error rates and decrease the risk of Type I error for this and for subsequent significance tests for main and interaction effects on score (Holm, 1979).

My original research plan had not included a repeated measures *t* test to investigate group-level score difference but instead a mixed-design ANOVA to compare exam scores (i.e., the dependent variable) between exam attempt (the within-subjects factor) between repeat examinees grouped by initial performance level (the between-subjects factor). I had planned to group repeat examinees into one of two groups based on initial performance: borderline failing examinees and clearly failing examinees. I had defined a borderline failing examinee as one whose initial exam score fell within one standard error of measurement (SEM) below the pass point and a clearly failing examinee's initial score fell more than one SEM below the pass point. However, after conducting the Rasch analyses to derive exam scores, I found that none of the scores of the repeat examinees in the study sample happened to fall within one SEM below the pass point. The borderline failing examinees in the dataset had either initially taken the exam prior to the five-year period of interest or did not repeat the exam within the five-year period. Therefore, no borderline failing examinees belonged to the sample of participants in this study. As a result, I replaced the mixed-design ANOVA with a repeated measures *t* test to investigate if overall exam scores differed significantly between initial and repeat exam attempts.

To still be able to evaluate repeat performance in relation to initial performance, I also looked at score difference for each individual in the study sample. I considered the standard errors (SEs) associated with both repeat examinees' initial scores and repeat scores. The Winsteps software program provides the SE associated with each Rasch score estimate (Linacre, 2009a). For dichotomous data, the SE associated with each Rasch score estimate is shown by the equation (2):

$$SE_{B_n} = \frac{1}{\sqrt{\sum_{i=1}^{L}(P_{ni})(1-P_{ni})}} \quad (2)$$

where

B_n = estimated level of ability of examinee *n*,

L = test length, and

P_{ni} = probability of examinee *n* answering item *i* correctly.

The SE associated with each score estimate can be used to define the confidence interval (CI) within which an examinee's true score would fall. First, I built 95% CIs of each initial score using the interval of initial score plus or minus 1.96 times the SE associated with that initial score. This allowed me to identify for how many and for which repeat examinees the pass point was within measurement error of initial score. Such repeat examinees now served as the borderline failing examinees, with the remaining examinees serving as those who clearly failed the exam. Next, I built 95% CIs of each repeat score using the interval of repeat score plus or minus 1.96 times the SE associated with that repeat score. For each repeat examinee, I then compared the initial score CI to the repeat score CI in order to identify overlapping CIs. Repeat examinees whose CIs did not overlap suggests that these examinees attained significantly different scores on initial versus repeat exam attempt. Comparing each repeat examinee's pair of

CIs allowed me to determine which repeat examinees earned significantly different scores on their repeat attempts and whether those score differences were gains or losses.

To answer Question 1(b), I applied the Board's criterion-referenced pass point to the overall exam scores. From there, I derived the rate of repeat examinees who passed the certification exam upon retesting. This pass rate indicates the proportion of repeat examinees who demonstrated an increase in knowledge sufficient enough to meet the minimum performance level needed for certification.

Question 1(c) focuses on whether any passing repeat examinees achieved repeat scores beyond measurement error of their initial scores. Higher scores within measurement error are to be expected. However, dramatic score gains might suggest score error, perhaps due to prior item exposure. As I mentioned in the previous chapter, it would be more reasonable to expect that examinees whose initial scores were within striking distance of the pass point, not far below it, are more likely to remediate their knowledge sufficiently enough to pass their repeat attempt. Therefore, identifying suspiciously high retest scores among passing repeat examinees remains important to investigating score errors and the quality of retest classifications.

To answer Question 1(c), I used results from both Questions 1(a) and 1(b). I had already evaluated the CIs for repeat examinees' initial and repeat scores, identifying those who had achieved score differences beyond measurement error. I had also identified which repeat examinees passed the certification exam upon retesting. I used those previous results to determine how many and which repeat examinees passed their repeat attempts with scores beyond their corresponding 95% CIs.

Among passing repeat examinees, any score gains beyond the 95% CI might suggest that they may have passed with the aid of a memory advantage from prior item exposure, thereby

warranting further investigation. After identifying passing individuals with exceedingly large score gains, I paid special attention to these cases throughout the remainder of my study to look for any supporting evidence of practice effects or memory advantages from prior item exposure. For example, if in a subsequent data analysis, I found that such an individual performed significantly better on common items compared to unique items and answered in a considerably shorter amount of time many common items correctly that they had previously gotten incorrect, then this result might suggest that the individual may have benefited from prior item exposure. If I later found that a passing repeat examinee with an exceeding score gain performed better on more cognitively complex items and not just on recall items, then this result might suggest truly remediated knowledge, thereby supporting his or her pass outcome.

2. Research Question 2

   Does examinee performance on different types of items lend support to the pass–fail determinations made about repeat examinees?

   a. How do reused exam items function with respect to distinguishing between different levels of competence among all examinees?
   b. Do repeat examinee subscores, on subtests of items grouped by content area and cognitive complexity level, differ significantly between initial and repeat exam attempt?

With Question 2(a), I was interested in how the reused items that repeat examinees saw more than once function across all examinees, passing or failing. To answer this research question, I first conducted a traditional item analysis on all scored items across all five exams using the initial or only response strings of all 864 examinees who had initially taken the exam within the five-year period of interest in this study. From there, I isolated the p-values and point-

biserial correlations for only the reused items to evaluate the quality of items that repeat examinees ever re-challenged.

Whereas the published guidelines on how to evaluate item difficulty and item discrimination indices vary, authors have agreed that evaluation criteria should primarily depend on the purpose of the exam (Downing, 2009a; Ebel & Frisbie, 1991; Haladyna, 2004; Schmeiser & Welch, 2006). Regarding p-value, moderately difficult items are generally the most informative on achievement tests, though including easier or more difficult items on an exam is sometimes necessary for content coverage (Downing, 2009a; Ebel & Frisbie, 1991; Haladyna, 2004). Schmeiser and Welch (2006) suggested a general guideline of inspecting any items with p-values less than .30. In their study on quality of multiple choice items on a summative assessment for undergraduate medical students, Malau-Aduli and Zimitat (2012) considered items with p-values between .20 and .80 to be acceptable. A certification handbook from the National Organization for Competency Assurance (NOCA) suggests an acceptable range of .33 to .92 for a criterion-referenced test (Dungan, 1996).

In this study, I classified items with p-values less than .30 as too difficult, between .30 to .40 as difficult, from .40 to .80 as ideal, and greater than .80 as easy. With the pass point of the certification exam in this study translating to a raw percent correct around 60% on each of the five exam forms, I first defined .40 to .80 as the ideal range for p-values. I classified items with p-values between .30 and .40 as difficult and items with p-values greater than .80 as easy. Compared to items in the ideal range, these difficult and easy items might not be as informative near the pass point, but their inclusion in scoring after the Board conducted item analyses suggests that these items were retained because they tested crucial knowledge. Lastly, I classified items with p-values less than .30 as too difficult. If the goal of a certification exam is

to identify minimum competence, then a lower bound that is above chance on an item with four response options is reasonable.

Regarding evaluation criteria for item discrimination, an item should have a positive discrimination value (Downing, 2009a). A negative discrimination index indicates that some item flaw is prompting more able examinees to answer the item incorrectly while less able examinees are answering the item correctly. That is, negative discrimination suggests that the item is flawed, thereby detracting from exam validity and reliability (Downing, 2009a; Ebel & Frisbie, 1991; Haladyna, 2004; Schmeiser & Welch, 2006). As for the threshold between a weak discrimination value and a strong value, researchers have used different thresholds. Schmeiser and Welch (2006) proposed flagging for review any items with discrimination values less than .20. In their study, Malau-Aduli and Zimitat (2012) defined a point-biserial correlation of .15 and above as discriminating and anything less than .15 as non-discriminating. The guideline from the NOCA handbook is a point-biserial correlation of .10 and higher as appropriate (Dungan, 1996).

When evaluating point-biserial correlations, I defined any items with point-biserial correlations of .15 or above as items with high discrimination and less than .15 as low discrimination, with the exception of items that also had p-values of .90 and higher. On an exam designed to assess minimum competence, there may be items that test important knowledge that almost all examinees answer correctly, which would result in low discrimination values for those items (Schmeiser & Welch, 2006). Accordingly, I categorized items with positive, low point-biserial correlations yet p-values of .90 or higher as also having high discrimination.

As I stated earlier in this chapter, board SMEs had already conducted an item analysis before scoring exams. However, they did this analysis for each exam administration separately.

Moreover, they may not have removed all items with poor statistical performance because of content coverage considerations. I therefore pooled all exam administrations for this item analysis for an enhanced view of item functioning. I focused this part of my study on identifying any reused items which were either very difficult among all certification examinees, not only those who failed and comprise the study sample, or which poorly discriminated between different levels of competence among all examinees. For any such flawed items that remained on subsequent forms of the exam, I further investigated those individual items to help uncover any issues with repeat examinees' scores.

To answer Question 2(b), I analyzed subscore gains and losses on different types of items. First, I categorized the exam items by content area and cognitive complexity. I identified four item types based on the content areas and cognitive complexity levels that the exam covers: basic biomedical science recall, basic biomedical science application, clinical science recall, and clinical science application.

In dividing the content domain by basic biomedical science and clinical science, the Board already classified individual exam items by those two major content areas. I used those existing classifications for this part of my study. Even though the Board further classified items using over 10 topic categories, I did not use these topic categories, because dividing the items across so many categories would have resulted in less reliable subscore estimates, making subscores less meaningful (Haladyna & Kramer, 2004; Monaghan, 2006; Puhan & Liang, 2011). Classifying items by the two major content areas instead helped mitigate this issue while still permitting deeper insight into repeat examinees' respective levels of content knowledge. It is important to note that any resulting subscores were nevertheless somewhat less stable compared to overall exam score estimates.

I identified the level of cognitive complexity for each exam item. To ensure that I identified them appropriately, I simplified the cognitive classifications. The item writers targeted cognitive processes such as recall, interpretation, and clinical judgment when developing the exam items. However, some authors have recommended classifying items as either "recall of an isolated fact" or "application of knowledge" as a more appropriate approach, arguing that it is challenging to conclusively determine which specific cognitive process examinees engage in when responding to an item because the thought process can vary between examinees (Case & Swanson, 2002; Clauser et al., 2006). Following such recommendations, I classified each item as either a *recall* item, in which the examinee has to remember or recognize a concept in order to answer the item, or as an *application* item, in which the examinee must recognize the content of the item, reach a conclusion, or select a course of action to answer the item. A psychometrician who has worked with the Board on their exam development for several years double-checked my cognitive classifications.

Next, I once again used Rasch analysis of the response data, this time to derive the repeat examinee subscores for each item type. Specifically, I used the Winsteps software program to analyze all examinees' responses on each subtest of related items and thereby estimate the repeat examinees' subscores. To ensure comparability of subscores across exam administrations, I anchored the items to their difficulty estimates from the five overall exam Rasch analyses that I had conducted to answer research Question 1 (Linacre, 2009a; Puhan & Liang, 2011). With four different item types in this part of my study, I used Winsteps to conduct four Rasch analyses for each of the five exam administrations.

Last, I conducted a two-way repeated measures ANOVA to compare repeat examinee subscores (the dependent variable) between subtest (a within-subjects factor) between exam

attempt (a within-subjects factor). This ANOVA allowed me to evaluate whether repeat examinees demonstrated on both attempts the same level of ability across the full range of content areas and skills represented on the exam. It also allowed evaluation as to whether any of the content area/cognitive complexity subtests lent better to demonstrated improvement on the retest. Again, I applied the Holm-Bonferroni method to adjust the alpha levels for this significance test (Holm, 1979).

My original plan to answer Question 2(a) involved a mixed-design ANOVA that included retest performance level as a between-subjects factor. Specifically, I had planned to group repeat examinees into one of four groups based on retest performance: clearly passing, borderline passing, borderline failing, and clearly failing. I had defined clearly passing as an examinee with a retest score falling more than one SEM above the pass point. A borderline passing examinee's score was within one SEM above the pass point. A borderline failing examinee was one whose retest score fell within one SEM below the pass point, and a clearly failing examinee's score fell more than one SEM below the pass point. However, after classifying the 62 repeat examinees into one of these four retest performance groups, I found that 43 repeat examinees fell within the clearly failing group, while only four, eight, and seven examinees fell into the clearly passing, borderline passing, and borderline failing groups, respectively. Because this would have resulted in heavily unequal cell sizes, consequently diminishing the meaning of any results from the ANOVA, I dropped retest performance level as a between-subjects factor. Conducting a two-way repeated measures ANOVA instead still facilitated investigation of group-level subscore differences by subtest of items grouped by content area and cognitive complexity.

3. Research Question 3

Do repeat examinee performances on items to which they have had prior exposure indicate any memory advantages or disadvantages from prior item exposure?

a. Do repeat examinees score differently on common items compared to unique items?
b. How do rates of response persistence and response change compare against changes in response time among all repeat examinees?
c. What are the results from 3(b) specifically among passing repeat examinees with score gains beyond measurement error, as identified in Question 1(c)?

The purpose of Question 3(a) was to determine whether repeat examinees score differently on common items compared to unique items. To answer the question, I first used Winsteps to estimate the repeat examinees' subscores on common items and on unique items for both initial and repeat exam attempts, again anchoring item difficulties to their benchmark scale difficulties (Linacre, 2009a; Puhan & Liang, 2011). Because some repeat examinees did not take the exam in consecutive years, not all repeat examinees in a given administration had prior exposure to the same set of items. This resulted in eight combinations of common and unique items that occurred among study participants. Therefore, I conducted the Winsteps analyses for each of the eight combinations, rather than simply conducting two analyses for each exam administration, and extracted from all the analyses the appropriate subscores for each participant. Again, it is important to note that any subscore estimates are less stable compared to overall exam score estimates because subscores are based on fewer items. Similarly, the common item subscores were generally somewhat less stable than the unique item subscores because most of the exam forms in this study generally contained fewer reused items than new items.

After arriving at the repeat examinees' subscores, I conducted a two-way repeated measures ANOVA to compare repeat examinee subscores (the dependent variable) by item type (common versus unique, a within-subjects factor) between exam attempt (a within-subjects factor). To adjust the alpha levels for this ANOVA, I again applied the Holm-Bonferroni method (Holm, 1979).

Originally, in place of this ANOVA, I had planned to conduct a repeated measures *t* test to compare average retest subscores on common items and on unique items. However, the ANOVA I ultimately conducted facilitated evaluation of any memory advantage from prior item exposure at a group-level by allowing comparison of improvement on common items versus improvement on unique items. If repeat examinees specifically improve on common items, this might suggest a memory advantage. Accompanying improvement on unique items would lend support to overall content knowledge remediation as the primary explanation for observed retest score gains. When investigating how repeat examinees performed on reused questions on a screening exam for physicians wanting to enter supervised practice, Wood (2009) had similarly conducted a two-way repeated measures ANOVA with exam attempt (Test 1 and Test 2) and question type (reused and non-reused) both as within-subject factors.

It remains important to ensure that each group of unique items that repeat examinees challenged between exam attempts was comparable in order for the comparison of unique item subscores by exam attempt to have much meaning. Though each group of unique items on initial and repeat attempts is comprised of different individual items, they are comparable in terms of content coverage and difficulty.

First, they cover the same parts of the content outline of the exam. The Board in this study used the same content outline for each exam form across the five-year period. Each group

of common items across two given exam forms covered the same sections of that content outline. In turn, each group of unique items between those two exam forms covered the same remaining portions of the content outline. For each of the eight combinations of unique items in this study, Table I displays the percentages of unique items in each content area, basic biomedical science and clinical science, on both initial and repeat attempt exam forms. In each instance, the percentages of unique items in each content area were similar.

To ensure that item difficulties for all common and unique items were on the same scale of measurement, I had anchored the items' difficulty values to their values from the overall exam analyses when conducting the Winsteps subscore analyses. Lastly, *z* tests indicated that the two unique item groups that each repeat examinee challenged were, on average, comparable in difficulty (see Table I). To conduct the *z* tests, I used the following equation (3):

$$z = \frac{D_I - D_R}{\sqrt{SE_I^2 + SE_R^2}} \quad (3)$$

where D_I and D_R are the mean estimated item difficulties for unique items on initial and repeat exam attempt, respectively, and SE_I and SE_R are the mean standard errors associated with these item difficulty estimates.

Table I

Comparability of Unique Items Between Initial and Repeat Exam Attempts

	Unique item combination							
Variable	1	2	3	4	5	6	7	8
n repeat examinees	24	5	3	1	2	14	1	12
Initial attempt								
n items	235	209	213	184	216	173	187	171
Item difficulty								
M	-0.05	-0.02	-0.03	0.09	-0.10	0.04	-0.06	-0.29
SD	1.29	1.31	1.17	1.20	1.17	1.27	1.24	1.17
SE	0.16	0.17	0.20	0.20	0.20	0.20	0.20	0.20
Item content area								
% items BBS	42.55%	30.62%	38.03%	44.02%	39.35%	33.53%	33.69%	32.16%
% items CS	57.45%	69.38%	61.97%	55.98%	60.65%	66.47%	66.31%	67.84%
Repeat attempt								
n items	195	173	215	168	222	154	191	191
Item difficulty								
M	0.01	0.10	0.00	-0.05	0.20	-0.13	0.25	0.22
SD	1.07	1.10	1.21	1.15	1.12	1.17	1.13	1.21
SE	0.20	0.19	0.20	0.19	0.19	0.20	0.19	0.19
Item content area								
% items BBS	43.08%	29.48%	39.07%	43.45%	39.64%	30.52%	32.98%	33.51%
% items CS	56.92%	70.52%	60.93%	56.55%	60.36%	69.48%	67.02%	66.49%
z test								
z	-0.23	-0.47	-0.11	0.51	-1.09	0.60	-1.12	-1.85
sig. (2-tailed)	.82	.64	.91	.61	.28	.55	.26	.06

Note. There were eight distinct combinations of unique items administered among the 62 repeat examinees in this study. BBS = Basic Biomedical Science; CS = Clinical Science.

In addition to comparing subscores at a group level through the repeated measures ANOVA, I also compared repeat attempt subscores at an individual level to better understand which repeat examinees may have experienced a memory advantage or disadvantage due to prior item exposure. I did this by conducting a *z* test for each repeat examinee, comparing retest subscore on common items and retest subscore on unique items. This allowed further investigation as to whether individual repeat examinees scored differently on common versus unique items and on which type of items they each performed better when retesting. To conduct these *z* tests, I used the following equation (4):

$$z = \frac{B_{nC} - B_{nU}}{\sqrt{SE_{nC}^2 + SE_{nU}^2}} \quad (4)$$

where

B_{nC} = retest subscore of examinee *n* on common items,

B_{nU} = retest subscore of examinee *n* on unique items,

SE_{nC} = the standard error associated with subscore of examinee *n* on common items, and

SE_{nU} = the standard error associated with subscore of examinee *n* on unique items.

As these individual *z* tests constituted multiple comparisons, I applied the Holm-Bonferroni Method to this set of individual comparisons to adjust minimum alpha levels and minimize Type I errors (Holm, 1979).

Question 3(b) focuses on how response selections and response times on common items during initial attempts compare to response selections and response times to those items during repeat attempts. First, across all common items, I ran crosstabs to find the frequency distributions of each possible response pattern, categorizing the patterns for each item as follows:

Correct–correct: The examinee selected the correct answer on both attempts.

Incorrect–correct: The examinee selected a distractor on the initial attempt, then the correct answer on the repeat attempt.

Correct–incorrect: The examinee selected the correct answer on the initial attempt, then a distractor on the repeat attempt.

Incorrect–same incorrect: The examinee selected the same distractor on both attempts.

Incorrect–different incorrect: The examinee selected a distractor on the initial attempt, then a different distractor on the repeat attempt.

Across all common items as a whole, I examined the rates of response persistence and response change against changes in response time across the common items as a whole. To accomplish this, I used the results from the preceding procedure to find the frequency distributions of each possible response pattern across all common items in the dataset. Next, I calculated the corresponding mean initial response time, mean repeat response time, and mean change in response time for each of the response patterns. Comparing response patterns with changes in response time across all of the common items helped determine whether the repeat examinees generally demonstrated increased content knowledge. For instance, higher rates of correct–correct and incorrect–correct response patterns, especially with somewhat shorter response times, suggest increased knowledge and possibly a memory advantage. High rates of correct–incorrect response changes or incorrect–incorrect patterns along with reduced response time indicate a memory disadvantage from prior item exposure. Fewer incorrect–correct changes than incorrect–same incorrect patterns highlight an inability to identify and address knowledge gaps. To provide a point of comparison for the common item response patterns and mean changes in response time, I also calculated the mean response times for correct and incorrect response selections on unique items.

Next, I conducted a similar procedure to compare response patterns against the mean changes in response time to evaluate whether repeat examinees succeeded or struggled with specific types of common items. Because I had already classified individual items by p-value, point-biserial correlation, content area, and cognitive complexity level from the analysis procedures for Questions 2(a) and 2(b), I was able to evaluate the response pattern and response time results for items grouped by these classifications. Similar to the response pattern analysis across all common items, I ran crosstabs to find the frequency distributions of each possible response pattern across all common items in each item group. For each response pattern, I then calculated the mean response times and mean changes in response time. To provide a point of comparison, I also calculated the mean response times for correct and incorrect response selections on unique items in each item group.

Again, the goal of comparing response patterns and response times by item group was to further explore whether different degrees of item quality or certain types of items related to indications of memory advantages or disadvantages. For example, on more difficult items, if many repeat examinees switched from a distractor to the correct answer while considerably reducing response time, then these repeat examinees may have unfairly benefitted from prior item exposure. That is, after initially taking the exam, examinees may have remembered and researched some of the notably difficult items, giving them an advantage when these items were reused on their repeat attempts. On the other hand, if many repeat examinees switched from correct answers to distractors on weakly discriminating common items, then poor item quality may have related to inability to rectify initial errors. Evaluating response pattern and response time results in relation to different item characteristics allowed a deeper look at item quality and the item characteristics that relate to the different response patterns that signify knowledge, false

knowledge building, knowledge remediation, and other memory advantages or disadvantages from prior item exposure.

Question 3(c) focuses specifically on the response patterns of repeat examinees who passed with score gains beyond measurement error, whom I had already identified when carrying out the procedure for Question 1(c). For each examinee, I ran a similar procedure twice to evaluate response patterns and their corresponding response times, once for patterns across all common items encountered and again for patterns by subtest of items grouped by content area and cognitive complexity. For each examinee, the procedure started with running crosstabs to find the frequency distributions of each possible response pattern, using the same response pattern classifications I listed above, across the common items he or she encountered. For each of the response patterns, I then calculated the examinee's corresponding mean change in response time. I used the results from these comparisons to look for indication of a memory advantage from prior item exposure. For instance, if the majority of an individual repeat examinee's response pattern changes were incorrect–correct and he or she considerably reduced response time with these changes, then this trend along with the magnitude of his or her score gain suggest a problematic memory advantage from prior item exposure. When conducting this procedure by content area/cognitive complexity subtest, I used the item classifications from my data analysis procedure in Question 2(b). I used the classifications from this set of subscore comparisons to help determine if these particular repeat examinees sufficiently remediated their knowledge gaps across all content areas and cognitive complexity levels reflected on the exam.

F. Summary of Research Method

In this chapter, I described my research approach. I described the repeat examinees whose performances I studied, the development of the certification exam forms used to assess

these examinees, the administration and scoring procedures for the exam, and my data analysis methods.

To accomplish the first part of my study, I compared overall score differences between initial and repeat exam attempts. I also examined the frequency with which score gains translated to passing the repeat exam attempt and determined if any such score gains were beyond measurement error. Next, I studied the performances on different types of items to assist interpretation of repeat examinees' score differences and pass–fail outcomes. This part of my investigation included comparing response patterns among all examinees, passing or failing, on initial exam attempts to gauge how well the reused exam items distinguished between different levels of competence. This part of the study also included comparing how repeat examinees performed between subtests of items grouped by content area and cognitive complexity between exam attempts. These comparisons helped provide deeper insight into the participants' respective levels of knowledge during initial and repeat testing. I focused the last part of my study on detecting any indication of prior item exposure on retest performances. This included comparing the participants' scores on common versus unique items. I also compared repeat examinees' changes in response pattern against changes in response time on common items when retesting. This, along with the results from the preceding analyses, allowed me to appraise any issues concerning the reuse of multiple choice items when assessing repeat examinees.

My methods, particular for Questions 2 and 3, generated an array of response patterns to interpret. Table II illustrates how I interpreted different types of results to make inferences regarding prior item exposure in terms of knowledge, knowledge remediation, false knowledge building, and inability to address knowledge gaps among the medical certification repeat examinees in this study.

Table II

Possible Inferences From Observed Repeat Examinee Subscores and Response Patterns

Variable	Result	Possible inference
Subscore on content area/cognitive complexity subtest	Increased score on both recall and application items	Knowledge remediation
	Increased score on recall items but not on application items	Memory advantage
	Increased score in both basic science and clinical science content areas	Knowledge remediation
	Increased score in one content area but not the other	Memory advantage
	Decreased score in one content area but not the other	Memory disadvantage
Subscore on common versus unique items	Increased score on common items and on unique items	Knowledge remediation
	Increased score on common items but significantly lower score on unique items	Memory advantage
	Decreased score on common items and comparable or increased score on unique items	Memory disadvantage
Response pattern and response time difference on common items	Correct–correct and similar or shorter response time	Persistent knowledge
	Correct–correct and longer response time	Persistent knowledge, increased carefulness or uncertainty on the repeat attempt
	Incorrect–same incorrect and similar or shorter response time	Persistent false knowledge
	Incorrect–same incorrect and longer response time	Persistent false knowledge, increased carefulness or uncertainty on the repeat attempt

Table II (*continued*)

Variable	Result	Possible inference
Response pattern and response time difference on common items	Incorrect–correct and similar or longer response time	Knowledge remediation
	Incorrect–correct and shorter response time	Memory advantage from prior item exposure
	Correct–incorrect and any response time	Uncertainty on the initial attempt, carelessness on the repeat attempt, or false recall
	Incorrect–different incorrect and similar or longer response time	Ability to recognize knowledge gap yet persistent lack of knowledge, misguided knowledge remediation, or eventual guessing on the repeat attempt
	Incorrect–different incorrect and shorter response time	Misguided knowledge remediation or guessing on the repeat attempt

IV. RESULTS

This chapter presents the results of my research. The results are organized in order of the research questions I posed to investigate how repeat examinees performed on a single medical certification specialty exam.

A. Research Question 1

Which repeat examinees' scores were most amenable to change between initial and repeat exam attempts, and to what extent did score differences indicate sufficiently remediated knowledge?

a. For repeat examinees who initially borderline failed and examinees who initially clearly failed the exam, do overall exam scores differ significantly between initial and repeat exam attempts?

b. What is the pass rate among repeat examinees?

c. Among passing repeat examinees, how many score gains are beyond measurement error?

To derive overall exam scores, I conducted a Rasch analysis for the benchmark exam and for each exam form during the five-year period of interest, equating the exams to the benchmark scale (see Appendix). Across the five Rasch analyses of interest, a total of 1,044 response strings across 924 individual examinees in the dataset were included. Common-item equating allowed examinee scores and item difficulty estimates of all five exams to be on the same scale of measurement. Table III displays a summary of all examinee scores and all examinee pass–fail rates, as well as numbers of scored items, reused items, and new items by year. Each year most examinees passed, with pass rates ranging from 60% to 69% and an unweighted mean pass rate of 66%. The pass rate for repeat examinees, including pre-Year 1 examinees not in the study

sample, ranged from less than 1% to 4%, which an unweighted mean pass rate of about 2%. Given that repeat examinees on average demonstrate lower performance than average examinees, it is unsurprising that the pass rate for repeat examinees was consistently considerably lower than that for initial examinees each year.

Table III

Summary of Overall Exam Performance for All Examinees by Year

Variable	Year 1	Year 2	Year 3	Year 4	Year 5
Examinees					
N	329	169	174	191	181
n fail (%)	101 (30.7%)	55 (32.5%)	69 (39.7%)	64 (33.5%)	57 (31.5%)
n pass (%)	228 (69.3%)	114 (67.5%)	105 (60.3%)	127 (66.5%)	124 (68.5%)
n initial pass (%)	219 (66.6%)	107 (63.3%)	104 (59.8%)	123 (64.4%)	119 (65.7%)
n repeat pass[a] (%)	9 (2.7%)	7 (4.1%)	1 (0.6%)	4 (2.1%)	5 (2.8%)
M score (logits)	0.95	0.88	0.82	0.92	0.91
SD score (logits)	0.49	0.50	0.53	0.53	0.52
Separation Reliability	.92	.92	.93	.92	.93
Exam items					
N administered	350	300	300	300	300
N scored	323	285	287	267	291
n reused (%)	—	87 (30.5%)	151 (52.6%)	188 (70.4%)	166 (57.0%)
n new (%)	—	198 (69.5%)	136 (47.4%)	79 (29.6%)	125 (43.0%)
Separation Reliability	.98	.96	.96	.96	.96

Note. This table summarizes the results of a total of 1,044 exam attempts across 924 individual examinees over the five-year period.
[a]Includes repeat examinees who initially took the exam prior to the five-year period and thus are not in the study sample.

Specifically regarding the sample of repeat examinees in this study, a summary of their overall exam scores on both initial and repeat exam attempts are in Table IV. As a whole, participants' total test scores were higher on repeat exam attempts (M = 0.41, SD = 0.39) than on initial exam attempts (M = 0.23, SD = 0.23). On average, participants improved their scores by 0.18 logits. The results from a repeated measures t test indicated that this average difference between initial score and repeat score was statistically significant, $t(61) = 4.80$, $p < .001$, d = 0.67.

The scatterplot in Figure 2 shows individual results among participants. Values on the identity line represent individuals with very minimal score differences between exam attempts. Values above the identity line represent individual score gains from initial to repeat exam attempt, and values below the identity line represent individual score losses. There was a strong, positive linear relationship between initial attempt score and repeat attempt score, $r(62) = .632$, $p < .001$. Of the 62 repeat examinees in the study sample, most (n = 46, 72.4%) improved their scores when they retook the certification exam. Review of the 95% CIs around repeat examinees' overall exam scores revealed that 63% of the repeat examinees achieved higher scores that were still within measurement error of initial score. About 11% of repeat examinees achieved statistically significant higher scores beyond measurement error. A little more than one fourth of repeat examinees experienced score losses. All observed score losses were within measurement error of initial and repeat scores, so lower repeat attempt scores were not statistically significantly different from initial scores, with 95% confidence.

Table IV

Descriptive Statistics for Repeat Examinee Overall Exam Scores

Exam attempt and examinee group	Overall exam score (logits)				
	M	*SD*	Mean *SE*	Minimum	Maximum
By initial performance group					
Initial attempt					
Borderline failing[a] (*n* = 6)	0.47	0.02	0.14	0.45	0.49
Clearly failing[b] (*n* = 56)	0.20	0.23	0.13	-0.72	0.46
Repeat attempt					
Borderline failing (*n* = 6)	0.67	0.50	0.13	-0.01	1.53
Clearly failing (*n* = 56)	0.38	0.37	0.13	-0.75	1.30
By retest pass–fail group					
Initial attempt					
Passing (*n* = 12)	0.33	0.16	0.13	-0.08	0.48
Failing (*n* = 50)	0.20	0.24	0.13	-0.72	0.49
Repeat attempt					
Passing (*n* = 12)	0.91	0.26	0.14	0.72	1.53
Failing (*n* = 50)	0.29	0.31	0.13	-0.75	0.67
All repeat examinees (*N* = 62)					
Initial attempt	0.23	0.23	0.13	-0.72	0.49
Repeat attempt	0.41	0.39	0.13	-0.75	1.53

[a]The borderline failing examinees performance group is comprised of repeat examinees whose 95% CI of initial score contained the exam pass point. [b]All other repeat examinees comprise the clearly failing performance group.

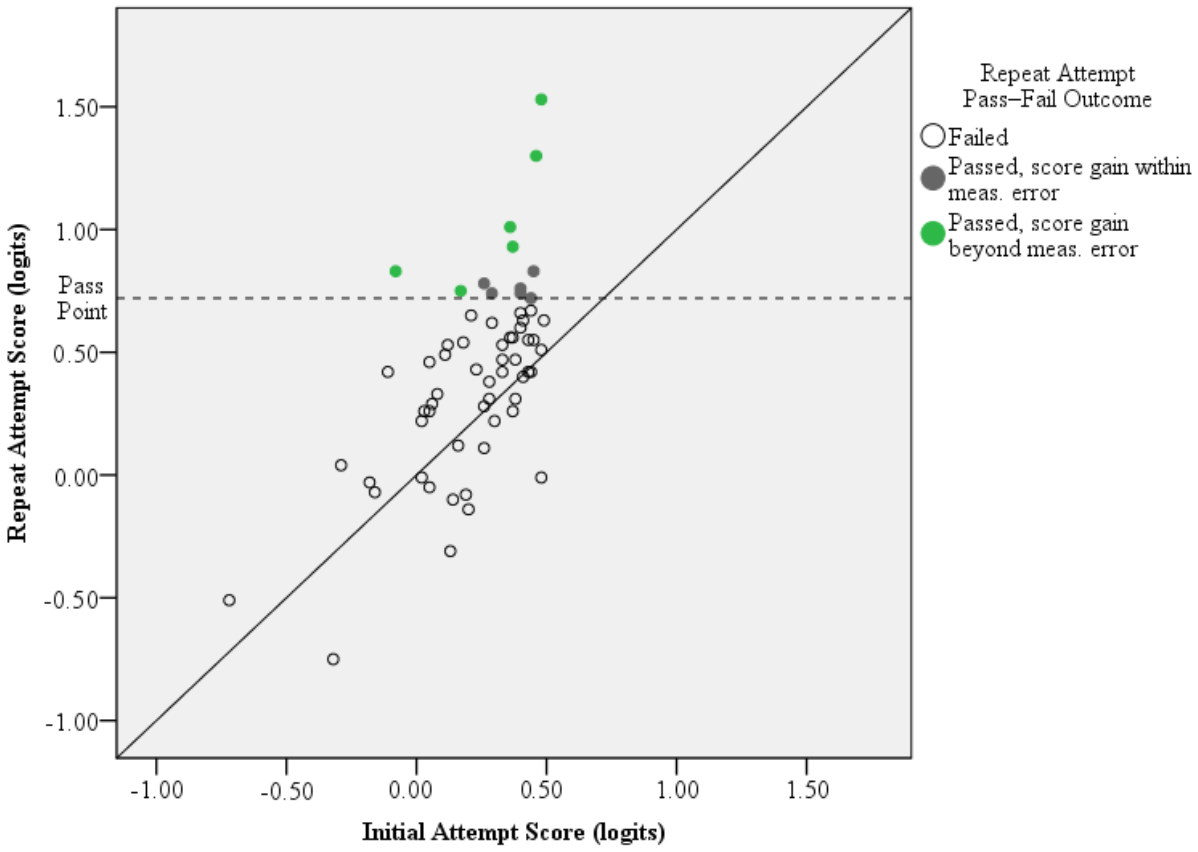

Figure 2. Comparison of overall exam scores between exam attempts.

Table V breaks down individual score differences by initial performance group, borderline failing or clearly failing. As a reminder, all participants whose 95% CI of initial score contained the exam pass point constituted the borderline failing performance group. Only six such repeat examinees fell into the borderline failing group. The 56 remaining examinees comprised the clearly failing performance group. Comparison of individual score differences by initial performance revealed that borderline failing and clearly failing repeat examinees experienced similar rates of score gains within measurement error of initial score. Both performance groups experienced at least one score gain beyond measurement error. As might be expected, a higher rate of score losses occurred among the participants who clearly failed their initial exam attempts. Again, all score losses were within measurement error.

Table V

Frequencies of Score Gains, Score Losses, and Pass–Fail Outcomes by Initial Performance Group

Performance result		Initial performance group: Borderline failing (*n* = 6)	Initial performance group: Clearly failing (*n* = 56)	All repeat examinees (*N* = 62)
Score gain	*n* within 95% CIs (%)	4 (66.7%)	35 (62.5%)	39 (62.9%)
	n beyond 95% CIs (%)	1 (16.7%)	6 (10.7%)	7 (11.3%)
Score loss	*n* within 95% CIs (%)	1 (16.7%)	15 (26.8%)	16 (25.8%)
	n beyond 95% CIs (%)	—	—	—
Repeat attempt outcome	*n* pass (%)	2 (33.3%)	10 (17.9%)	12 (19.4%)
	n fail (%)	4 (66.7%)	46 (82.1%)	50 (80.6%)

Note. The borderline failing examinees performance group is comprised of repeat examinees whose 95% CI of initial score contained the exam pass point. All other repeat examinees comprise the clearly failing performance group.

Even though participants generally improved their scores when retaking the certification exam, most did not improve enough to pass repeat exam attempts, as shown in the bottom portion of Table V. Only 19% of repeat examinees in this study passed their repeat attempt. The individual repeat examinees who did pass were among both the borderline failing and clearly failing initial performance groups. However, as a group, they had earned a higher mean exam score on initial attempt compared to those who ended up failing (see Table IV).

About four-fifths of repeat examinees failed to demonstrate sufficiently remediated knowledge when retesting. As a group, the participants who had borderline failed their initial attempts had earned a higher mean exam score on repeat attempt compared to those who had

clearly failed their initial attempts (see Table IV). However, only two of the six borderline failing examinees ended up passing when they repeated the exam. Therefore, even some of the participants with comparatively more bridgeable knowledge gaps failed to demonstrate sufficiently remediated knowledge when retaking the certification exam.

Of the 12 passing repeat examinees, six examinees passed with a repeat attempt score beyond their respective 95% CIs. These individuals are denoted with solid green dots in Figure 2. Later in this chapter, I provide results from analysis of their response patterns in order to contextualize their dramatic score gains in terms of knowledge remediation and memory advantages from prior item exposure.

B. Research Question 2

Does examinee performance on different types of items lend support to the pass–fail determinations made about repeat examinees?

a. How do reused exam items function with respect to distinguishing between different levels of competence among all examinees?

b. Do repeat examinee subscores, on subtests of items grouped by content area and cognitive complexity level, differ significantly between initial and repeat exam attempt?

1. Item analysis

Using the response strings from the initial or only exam attempts by all 864 examinees who initially took the certification exam during the five-year period covered in this study, I conducted a traditional item analysis on all 868 items scored over the five-year period. After conducting the item analysis, I isolated the results of the 358 reused items in the dataset in order to evaluate the quality of reused items that the study participants challenged more than

once. Table VI displays a summary of the results from the item analysis, both by content area/cognitive complexity subtest and across all reused items in the dataset.

Concerning the reused items as a whole, the majority of reused items had p-values within the ideal range of .40 to .80 (n = 261, 72.9%). Therefore, most items that participants repeatedly challenged were of moderate difficulty among all initial certification examinees. Similarly, the majority of reused items had positive point-biserial correlations for the correct answer (n = 349, 97.5%). That is, most items that participants re-challenged had satisfactorily discriminated between more able and less able examinees, thereby lending support to the pass–fail determinations made about repeat examinees.

On average, all subtests of items grouped by content area and cognitive complexity level were comparable in item difficulty, with mean p-values ranging from .65 to .69. For each subtest, the majority of reused items had p-values in the ideal range. The most difficult reused items tended to be clinical science recall items. Regarding item discrimination, basic biomedical science items were, on average, more discriminating than clinical science items. This result is not unexpected. Clinical science knowledge is more frequently reinforced while working in the clinical setting, whereas basic biomedical science knowledge encompasses the expertise needed to work with unfamiliar or extraordinary medical cases (Cianciolo, Williams, Klamen, & Roberts, 2013). Therefore, study participants may have been more likely to answer clinical science items correctly. For each subtest, the majority of reused items had high point-biserial correlations.

Table VI

Summary of Item Analysis Results for Reused Items

Index	Basic Biomedical Science		Clinical Science		All reused items
	Recall	Application	Recall	Application	
n items	94	25	151	88	358
Item difficulty					
Mean p-value (*SD*)	.67 (.15)	.69 (.15)	.65 (.18)	.68 (.16)	.66 (.16)
Minimum p-value	.26	.43	.17	.20	.17
Maximum p-value	.96	.90	.93	.94	.96
p-value frequencies					
Too difficult (< .30)	1 (1.1%)	–	6 (4.0%)	1 (1.1%)	8 (2.2%)
Difficult (.30 ≤ p-value < .40)	4 (4.3%)	–	8 (5.3%)	2 (2.3%)	14 (3.9%)
Ideal (.40 ≤ p-value ≤ .80)	72 (76.6%)	18 (72.0%)	102 (67.5%)	69 (78.4%)	261 (72.9%)
Easy (> .80)	17 (18.1%)	7 (28.0%)	35 (23.2%)	16 (18.2%)	75 (20.9%)
Item discrimination					
Mean r_{pbis} *(SD)*	.21 (.10)	.23 (.09)	.16 (.09)	.14 (.09)	.18 (.10)
Minimum r_{pbis}	-.07	.07	-.08	-.07	-.08
Maximum r_{pbis}	.55	.50	.38	.38	.55
r_{pbis} frequencies					
Negative	1 (1.1%)	–	4 (2.6%)	4 (4.5%)	9 (2.5%)
Low ($0 \le r_{pbis} < .15$)[a]	22 (23.4%)	3 (12.0%)	55 (36.4%)	33 (37.5%)	113 (31.6%)
High (≥ .15)	71 (75.5%)	22 (88.0%)	92 (60.9%)	51 (58.0%)	236 (65.9%)

Note. The item analysis was conducted using all scored items and all initial responses by all 864 examinees who initially took the exam within the five-year period of interest.

[a]Excludes items with p-values greater than .90. Items with p-values greater than .90 and any positive point-biserial correlation were classified as High.

Though the reused items were generally acceptable in terms of difficulty and discrimination, some items functioned poorly across all certification examinees, including non-repeating examinees. About 22 reused items (6.1%) may have been difficult or too difficult for the purposes of the exam. Of these 22 items, 14 items had at least one distractor with a p-value of .10 or greater and a positive point-biserial correlation. That is, the most difficult reused items had distractors that appealed to a considerable proportion of all initial examinees, including some of the more able examinees. Similarly, all negatively discriminating reused items and 37 weakly discriminating reused items had at least one such appealing distractor. These findings altogether indicate that even though most reused items were acceptable in terms of difficulty and discrimination, study participants re-challenged a number of possibly flawed, ambiguous items.

The reuse of poorly functioning items may have been due to a few reasons. First, SMEs may have deemed these items appropriate from a content standpoint and decided to reuse them on a subsequent exam form. That is, SMEs determined that such items were appropriately written and tested important, relevant medical content despite poor item statistics. Another possibility is that the Board had to reuse items with less desirable statistics in order to cover the content outline or increase the number of common items available for equating. Lastly, it is important to note that I conducted this item analysis across all years in the study, whereas the Board originally evaluated item analysis results separately for each exam administration. Items that functioned poorly in this study's item analysis may have functioned better in individual exam administrations.

To look at how the reuse of flawed items may have impacted participants' repeat attempt outcomes, I regenerated repeat attempt scores with all negatively discriminating, difficult, and too difficult reused items removed from scoring. Among the 62 participants, the outcomes of

three participants (4.84%) changed from fail to pass, and all other pass–fail outcomes remained the same. All three participants initially tested in Year 1 and repeated the exam in Year 2. Prior to item removal, their retest scores had fallen just below the pass point. Upon item removal, each of their scores increased by 0.12 logits to fall just above the pass point. Therefore, those problematic items did influence examinees' exam results, although only a few examinees were affected.

To further understand the implications of these findings for repeat examinee performance, when I provide the results for Question 3(b) later in this chapter, I include a comparison of rates of response persistence and of response change for the different levels of item difficulty and item discrimination among reused items. The goal of this was a deeper look into how item quality might relate to different repeat performance behaviors such as remediating knowledge, building false knowledge, and benefitting from memory advantages due to prior item exposure.

2. Content area/cognitive complexity subtests

Table VII summarizes participant subscores by content area/cognitive complexity subtest by year by exam attempt. As a reminder, the mean p-values for all four subtests were comparable (see Table VI). Of the 12 passing repeat examinees, four passed in Year 2, four in Year 4, and four in Year 5. Therefore, any specific distribution of items by subtest does not appear connected to a higher pass rate among repeat examinees.

Table VII

Content Area/Cognitive Complexity Subtest Performances for Repeat Examinees by Year (N = 62)

Variable	Year 1	Year 2	Year 3	Year 4	Year 5
			Item distributions		
N scored items	323	285	287	267	291
n scored items (% of *N* scored items)					
BBS Recall	102 (31.6%)	79 (27.7%)	86 (30.0%)	68 (25.5%)	72 (24.7%)
BBS Application	15 (4.6%)	22 (7.7%)	18 (6.3%)	25 (9.4%)	32 (11.0%)
CS Recall	133 (41.2%)	119 (41.8%)	103 (35.9%)	100 (37.5%)	99 (34.0%)
CS Application	73 (22.6%)	65 (22.8%)	80 (27.9%)	74 (27.7%)	88 (30.2%)
			Person separation reliability		
BBS Recall	.80	.84	.86	.81	.84
BBS Application	.32	.47	.45	.56	.62
CS Recall	.71	.77	.78	.79	.81
CS Application	.49	.66	.69	.67	.75
			Initial attempts		
n examinees	29	6	15	12	—
BBS Recall					
M subscore (*SD*)	-0.16 (0.42)	-0.01 (0.51)	-0.14 (0.33)	0.24 (0.39)	—
Mean *SE*	0.23	0.26	0.25	0.28	—
BBS Application					
M subscore (*SD*)	0.61 (0.76)	-0.13 (0.56)	0.29 (0.54)	0.66 (0.60)	—
Mean *SE*	0.58	0.50	0.53	0.48	—
CS Recall					
M subscore (*SD*)	0.30 (0.31)	0.17 (0.23)	0.17 (0.32)	0.20 (0.22)	—
Mean *SE*	0.20	0.20	0.22	0.22	—
CS Application					
M subscore (*SD*)	0.74 (0.27)	0.39 (0.50)	0.34 (0.26)	0.67 (0.26)	—
Mean *SE*	0.27	0.28	0.25	0.26	—

Table VII *(continued)*

Variable	Year 1	Year 2	Year 3	Year 4	Year 5
		Repeat attempts			
n examinees	—	24	8	15	15
n passing	—	4	0	4	4
% of *N* passing (12)	—	33.3%	0.0%	33.3%	33.3%
BBS Recall					
M subscore (*SD*)	—	0.16 (0.68)	-0.42 (0.48)	0.41 (0.69)	0.07 (0.67)
Mean *SE*	—	0.27	0.26	0.28	0.28
BBS Application					
M subscore (*SD*)	—	0.41 (0.66)	0.14 (0.44)	0.64 (0.70)	0.59 (0.89)
Mean *SE*	—	0.50	0.53	0.48	0.41
CS Recall					
M subscore (*SD*)	—	0.40 (0.30)	0.08 (0.41)	0.26 (0.37)	0.47 (0.47)
Mean *SE*	—	0.20	0.22	0.22	0.22
CS Application					
M subscore (*SD*)	—	0.65 (0.41)	0.38 (0.40)	0.66 (0.49)	0.82 (0.37)
Mean *SE*	—	0.28	0.25	0.26	0.24

Note. All subscore estimates are in logits. BBS = basic biomedical science; CS = clinical science.

Table VIII displays descriptive statistics for subscores across all participants. After normality checks showed normally distributed residuals, a two-way repeated measures ANOVA with a Huynh-Feldt correction revealed a significant main effect for subtest, $F(2.56, 156.16) = 42.52$, $p < .001$, $\eta_p^2 = .41$. Regardless of exam attempt, repeat examinees scored significantly differently across the content area/cognitive complexity subtests. Post hoc tests with a Bonferroni correction for the six pairwise comparisons revealed that participants demonstrated significantly lower ability on both recall subtests, basic biomedical science and clinical science, than on both application subtests. Basic biomedical science recall subscores were an average of 0.26 logits lower than clinical science recall subscores ($p < .001$). Similarly, participants scored an average of 0.16 logits lower on basic biomedical application rather than on clinical science application, though this difference was not statistically significant ($p = .079$, *ns*). They scored an average of 0.44 logits lower on basic biomedical recall items than on basic biomedical application items ($p < .001$) and an average of 0.34 logits lower on clinical science recall items than on clinical science application items ($p < .001$). Therefore, repeat examinees tended to perform better on application items and on clinical science items. For comparison, a one-way repeated measures ANOVA revealed that one-time examinees also performed significantly better on application items than on recall items, yet significantly worse on clinical science items than on basic biomedical science items.

Among the repeat examinees, there was no significant main effect for exam attempt on subscore, $F(1, 61) = 3.65$, $p = .122$, *ns*. Therefore, repeat examinee subtest subscores did not significantly differ across exam attempts regardless of type of subtest.

Though mean subscores increased by varying amounts for all four subtests, type of subtest remained unrelated to mean subscore gains. The ANOVA with a Huynh-Feldt correction

showed no significant interaction effect between exam attempt and subtest on subscore, $F(2.30, 140.42) = 1.22$, $p = .301$, *ns*. That is, mean subscore gains were not significantly different by content area/cognitive complexity subtest. Even though repeat examinees performed differently across the content area/cognitive complexity subtests, none of the subtests appeared to significantly lend better to score gains or losses.

Table VIII

Content Area/Cognitive Complexity Subtest Performances Across All Repeat Examinees (N = 62)

	Subcores (logits)				
Exam attempt and subtest	*M*	*SD*	Mean *SE*	Minimum	Maximum
Initial attempt					
Basic Biomedical Science					
Recall	-0.063	0.420	0.248	-1.16	0.84
Application	0.471	0.693	0.541	-1.17	2.12
Clinical Science					
Recall	0.235	0.290	0.209	-0.64	0.87
Application	0.598	0.336	0.264	-0.45	1.22
Repeat attempt					
Basic Biomedical Science					
Recall	0.126	0.690	0.272	-2.05	2.09
Application	0.474	0.711	0.478	-2.02	1.99
Clinical Science					
Recall	0.341	0.388	0.215	-0.93	1.16
Application	0.659	0.432	0.263	-0.37	1.53

The goal of investigating how repeat examinees performed across items from different content areas and of differing cognitive complexity through these significance tests was to yield greater insight into the participants' respective levels of knowledge during both exam attempts. I revisit performances by content area/cognitive complexity subtest later in this chapter when I present response pattern analysis results for Question 3(b).

C. Research Question 3

Do repeat examinee performances on items to which they have had prior exposure indicate any memory advantages or disadvantages from prior item exposure?

a. Do repeat examinees score differently on common items compared to unique items?

b. How do rates of response persistence and response change compare against changes in response time among all repeat examinees?

c. What are the results from 3(b) specifically among passing repeat examinees with score gains beyond measurement error, as identified in Question 1(c)?

1. Comparison of common item and unique item subscores

On average, repeat examinees re-challenged about 35% of all scored items during retesting (*SD* = 5%). The percentage of common items to all scored items that repeat examinees saw during retesting ranged from 24% to 42%. No specific percentage of common items appears to connect to a higher pass rate among repeat examinees (see Table IX).

There was no significant linear relationship between proportion of common items on repeat attempt and overall exam score difference between initial and repeat attempts, $r(62) = .093$, $p = .469$, *ns*. As a result, I excluded proportion of common items from the group-level significance test comparing subscore by item type (common or unique) by exam attempt.

Table IX

Number of Passing Repeat Examinees by Percentage of Common Items on Repeat Attempt

% Common items on repeat attempt	*n Passing examinees*	*% Passing examinees*
23.7%	1	8.3%
31.6%	4	33.3%
34.4%	3	25.0%
42.3%	4	33.3%

Table X presents descriptive statistics for common item and unique item subscores among repeat examinees, both grouped by pass–fail status after repeat attempt and across all repeat examinees. All assumptions were met for the ANOVA to compare average subscore by item type by exam attempt. There was a significant main effect for exam attempt, $F(1, 61) = 22.94$, $p < .001$, $\eta_p^2 = .27$. Regardless of item type, repeat examinees generally earned significantly higher subscores on their repeat attempt ($M = 0.40$ logits) than on their initial attempt ($M = 0.21$ logits). There was also a significant main effect for item type, $F(1, 61) = 7.28$, $p = .027$, $\eta_p^2 = .11$. As a group, participants scored an average of 0.11 logits higher on unique items compared to common items regardless of exam attempt. There was a significant effect for the interaction between exam attempt and item type, $F(1, 61) = 26.04$, $p < .001$, $\eta_p^2 = .30$. As a group, repeat examinees experienced greater average score gains on common items than on items to which they had no prior exposure, signifying a possible memory advantage.

Table X

Repeat Examinee Common Item and Unique Item Performance by Pass–Fail Group

	Initial attempt			Repeat attempt			Change		
Subscore (logits)	Common items	Unique items	Total test	Common items	Unique items	Total test	Common items	Unique items	Total test
				Passing examinees (n = 12)					
M	0.30	0.36	0.33	0.91	0.91	0.91	0.61	0.55	0.58
SD	0.32	0.20	0.16	0.31	0.30	0.26	0.36	0.33	0.24
Mean *SE*	0.22	0.16	0.13	0.23	0.18	0.14	—	—	—
Minimum	-0.30	-0.01	-0.08	0.58	0.58	0.72	0.00	-0.15	0.28
Maximum	0.73	0.73	0.48	1.44	1.55	1.53	1.12	1.00	1.05
				Failing examinees (n = 50)					
M	0.03	0.32	0.20	0.29	0.26	0.29	0.26	-0.07	0.09
SD	0.37	0.29	0.24	0.42	0.32	0.31	0.33	0.29	0.23
Mean *SE*	0.23	0.16	0.13	0.23	0.17	0.13	—	—	—
Minimum	-0.91	-0.51	-0.72	-0.87	-0.79	-0.75	-0.70	-1.15	-0.49
Maximum	0.66	1.15	0.49	1.05	0.70	0.67	1.07	0.48	0.53
				All repeat examinees (N = 62)					
M	0.09	0.33	0.23	0.41	0.38	0.41	0.32	0.05	0.18
SD	0.38	0.27	0.23	0.47	0.41	0.39	0.36	0.39	0.30
Mean *SE*	0.23	0.16	0.13	0.23	0.17	0.13	—	—	—
Minimum	-0.91	-0.51	-0.72	-0.87	-0.79	-0.75	-0.70	-1.15	-0.49
Maximum	0.73	1.15	0.49	1.44	1.55	1.53	1.12	1.00	1.05

Though repeat examinees may have generally experienced a memory advantage due to prior item exposure, such an advantage did not appear sufficient to achieving a passing score. As indicated in the right-hand portion of Table X, both failing and passing repeat examinees experienced a positive mean change in common item subscore. However, the mean subscore changes for both common and unique items are somewhat comparable among passing examinees, yet more discrepant among failing examinees. Moreover, comparison of subscores for each examinee by pass–fail outcome indicates that better or worse performance on common items did not directly relate to passing or failing repeat attempts (see Table XI). These results suggest that passing examinees also had to improve their overall content knowledge, not solely their knowledge of common item content, to attain their passing scores.

Table XI

Retest Common and Unique Item Subscore Comparisons by Pass–Fail Group

	Common > Unique		Common < Unique	
Pass–fail group	*n* examinees	*%* examinees	*n* examinees	*%* examinees
Pass (*n* = 12)	5	41.7%	7	58.3%
Score gain within 95% CIs (*n* = 6)	3	50.0%	3	50.0%
Score gain beyond 95% CIs (*n* = 6)	2	33.3%	4	66.7%
Fail (*n* = 50)	27	54.0%	23	46.0%

Inspection of individual subscore differences bolsters this finding. To evaluate individual-level score differences between common item and unique item subscores during repeat attempt, I conducted a *z* test between retest common item subscore and retest unique item subscore for each repeat examinee. The results revealed that no individual examinee performed significantly differently on common items than on unique items during repeat attempt, after using the Holm-Bonferroni method to correct for multiple comparisons (Holm, 1979).

The significant attempt-item type interaction from the ANOVA suggests that repeat examinees' retest common item subscores generally reflected a memory advantage. At the same time, on an individual level, retest common item and unique item subscores were not significantly different. This signifies that no passing individual performed significantly better on common items than on unique items when retesting, thereby lending support to their passing outcomes. Conversely, no failing individual performed significantly more poorly on common items than on unique items. Consequences of any memory disadvantages due to prior item exposure therefore appear limited, bolstering individual fail outcomes.

2. Common item response patterns with mean changes in response time

Results from the preceding analyses suggest a potential memory advantage due to prior item exposure. Even so, most repeat examinees were unable to achieve a passing score when retaking the exam. To help illuminate these concurrent findings, I present in this part of the chapter a closer look at how participants responded on common items on both initial and repeat exam attempts. First, I present the response patterns and response times across common items as a whole. To refine these general observations, over the four subsequent sections I hone in on the results for different groups of items. Specifically, I present pattern and time results on items grouped by their difficulty indices from Question 2(a). Following that are the results on

items grouped by their discrimination indices. Next are the results on items grouped by their content area/cognitive complexity level classifications from Question 2(b). Last, among only participants who passed with score gains beyond measurement error, I present the results across all common items as well as by content area/cognitive complexity subtest. Within each section, I describe response persistence patterns first and response change patterns second.

a. Across common items as a whole

Figure 3 displays the distribution of response patterns across all common items by pattern type. The most frequent response pattern was correct–correct, and the second most frequent pattern was incorrect–same incorrect. Correct–correct and incorrect–correct response patterns together accounted for 56% of observed patterns. In other words, both persistent knowledge and persistent false knowledge accounted for most common item responses. As for rates of response change, participants remediated initially incorrect responses about 19% of the time. However, not all response changes signified knowledge remediation or an advantage from prior item exposure, with correct–incorrect and incorrect–different incorrect patterns occurring about 13% and 12% of the time, respectively.

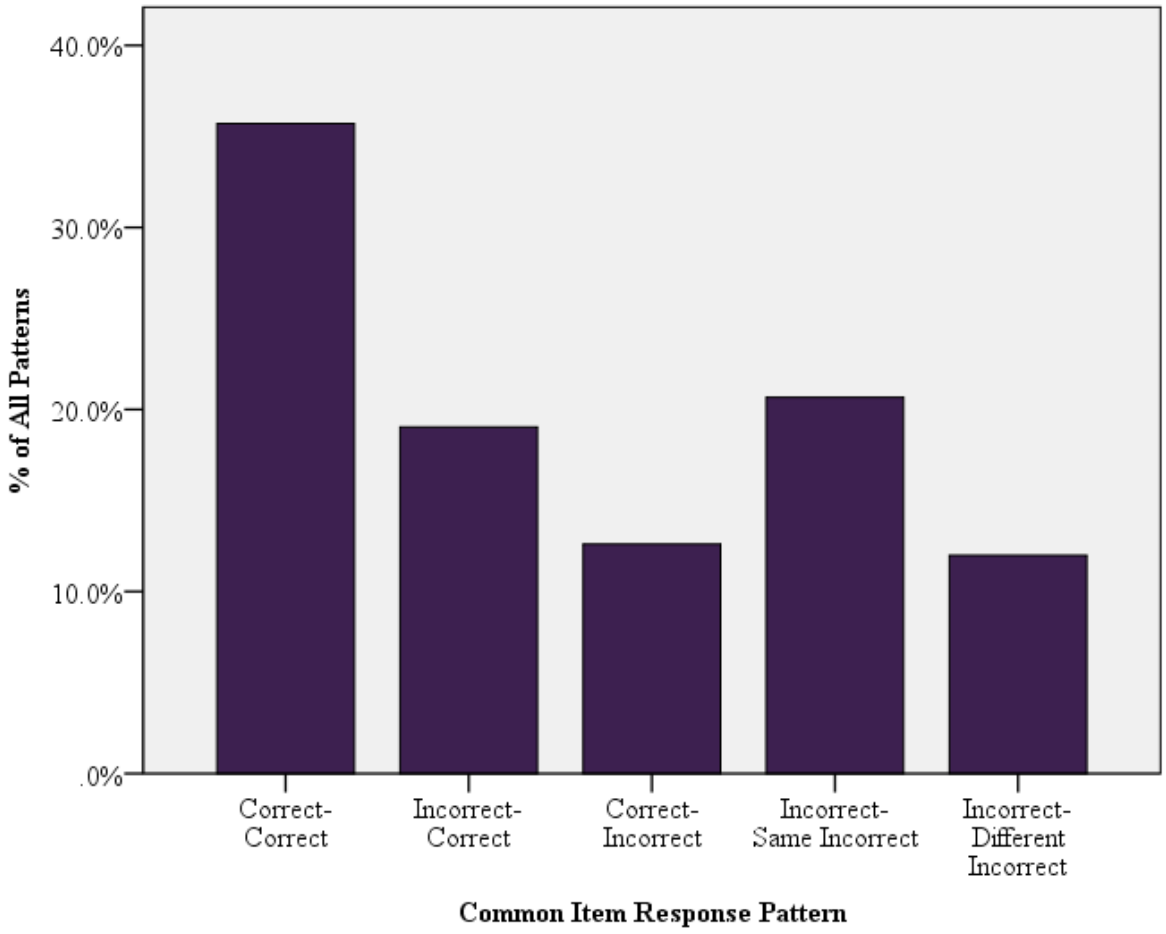

Figure 3. Common item response patterns across all items.

Table XII displays response patterns on common and unique items along with mean response times across exam attempts. All assumptions were met for a three-way repeated measures ANOVA to compare average response time by exam attempt by item type (common or unique) by item score (correct or incorrect). There was a significant main effect for item score, $F(1, 61) = 208.16$, $p < .001$, $\eta_p^2 = .77$. On average, participants took significantly less time to answer an item correctly ($M = 57.75$ s) than incorrectly ($M = 72.90$ s) regardless of exam attempt or item type. The main effect of exam attempt was also significant, $F(1, 61) = 10.66$, $p = .002$, $\eta_p^2 = .15$. Regardless of item type or item score, item response times were higher on repeat attempts ($M = 67.36$ s) than on initial attempts ($M = 63.29$ s). The main effect of item type was significant as well, $F(1, 61) = 26.11$, $p < .001$, $\eta_p^2 = .30$. On average, participants spent less time responding to unique items ($M = 64.17$ s) than to common items ($M = 66.48$ s).

The interaction between exam attempt and item type was significant, $F(1, 61) = 20.57$, $p < .001$, $\eta_p^2 = .25$. Surprisingly, post hoc tests revealed that response times between unique and common items significantly differed only during initial exam attempts. Despite having prior exposure to the common items, participants' response times were similar between common and unique items during repeat exam attempts.

Table XII

All Item Response Patterns With Mean Response Times (N = 30,209)

	Frequency		Response time (seconds)					
			Initial attempt		Repeat attempt		Change[a]	
Item type and response pattern	*n*	%	*M*	*SD*	*M*	*SD*	*M*	*SD*
Common item (*n* = 6,005)								
Correct–Correct	2,144	35.7%	56.19	47.12	55.47	50.89	-0.71	51.35
Incorrect–Correct	1,143	19.0%	76.27	62.72	69.10	60.09	-7.17	72.99
Correct–Incorrect	757	12.6%	70.20	51.02	80.33	59.77	10.13	64.31
Incorrect–Same Incorrect	1,241	20.7%	66.25	48.42	66.10	49.71	-0.15	55.54
Incorrect–Different Incorrect	720	12.0%	76.33	56.16	78.79	57.68	2.46	62.96
Unique item (*n* = 24,204)								
Correct	13,657	56.4%	51.61	44.51	59.06	53.24	—	—
Incorrect	10,547	43.6%	67.75	55.64	74.83	59.35	—	—

[a]Change in response time was calculated for each individual item. Thus, these values are blank for unique items.

To evaluate if response times differed between exam attempts between common item response patterns, I conducted a one-way ANOVA. First, I carried out a log transformation so that response times would follow a normal distribution. Because the assumption of homogeneity of variance was not met for the transformed data, I used *Welch's F* test to compare transformed response times between item groups based on both response pattern and attempt (e.g., correct–correct/initial, correct–correct/repeat). Overall, the logs of response times were significantly different between exam attempts between common item response patterns, $Welch's\ F(9, 4072.20) = 56.14$, $p < .001$.

Repeat examinees not only most frequently selected the same response options across exam attempts but also did so in roughly the same amount of time across attempts, as indicated by post hoc tests using the Games-Howell procedure. When providing the same response, participants may have therefore engaged in similar reasoning processes on both attempts. As correct–correct was the most frequent pattern, participants most frequently demonstrated persistent knowledge on common items. However, participants also demonstrated persistent false knowledge at a substantial rate, with incorrect–same incorrect being the second most frequent pattern.

Post hoc tests also revealed that the mean times for the incorrect–same incorrect response pattern were significantly longer than those for the correct–correct pattern on both attempts ($p < .001$). At the same time, they were significantly shorter than response times for incorrect–different incorrect responses on both attempts ($p < .001$). They were also shorter than the response times for the correct–incorrect pattern, though this difference was significant only on the repeat attempt ($p = .775$ and $p < .001$, respectively). On initial attempts only, they were significantly shorter than initially incorrect responses that were later remediated ($p = .002$).

Altogether, these results suggest less certainty with incorrect–same incorrect responses compared to correct–correct responses, yet more certainty with repeatedly selected distractors than with those they selected once.

Participants successfully remediated initially incorrect responses almost one fifth of the time. The shorter response time accompanying this change was significant ($p < .001$), reinforcing the possibility of a memory advantage from prior item exposure. Even so, the mean time for a correct response on a second attempt after initially answering incorrectly ($M = 69.10$ s) was significantly longer than mean times for correct–correct responses on initial attempts ($M = 56.19$ s, $p < .001$) and on repeat attempts ($M = 55.47$ s, $p < .001$). Moreover, the mean time for a remediated response was also longer than the mean response time for correct responses on unique items on initial attempts ($M = 51.51$ s) or repeat attempts ($M = 59.06$ s). Perhaps this additional time to remediate initial errors reflected an effort to retrieve the correct answer from whatever content knowledge participants accrued when preparing to retest.

On items with correct–incorrect responses, mean response time increased between exam attempts, though this difference was not significant ($p = .069$). However, mean response time when initially answering these items correctly (70.20 s) was significantly longer than when initially answering correctly on items with correct–correct patterns (56.19 s, $p < .001$), as well as longer than initially correct responses on unique items (55.14 s). Therefore, in changing from the correct answer to a distractor, repeat examinees may have been acting on doubt of their initially correct responses.

Incorrect–different incorrect responses comprised the smallest, yet still considerable, proportion of common item responses. Corresponding mean response times were comparable across exam attempts. At the same time, they were generally significantly longer compared to

mean response times for other response patterns. Therefore, incorrect–different incorrect responses might largely suggest engagement in erroneous reasoning and remediation.

b. By item difficulty index

In this section of the chapter, I present response patterns and their corresponding response times on items grouped by the levels of difficulty I formed during the item analysis for Question 2(a). As a reminder, I classified items with p-values less than .30 as too difficult, between .30 to .40 as difficult, from .40 to .80 as ideal, and greater than .80 as easy. Figure 4 shows that common item response pattern rates varied by and within item p-value group. On average, items on which repeat examinees demonstrated persistent knowledge (correct–correct) were the easiest, whereas items on which they demonstrated false knowledge or lack of knowledge (incorrect–same incorrect and incorrect–different incorrect) were the most difficult (see Table XIII). Interestingly, the mean p-value of the items they erroneously rectified was similar to that of the items they properly rectified.

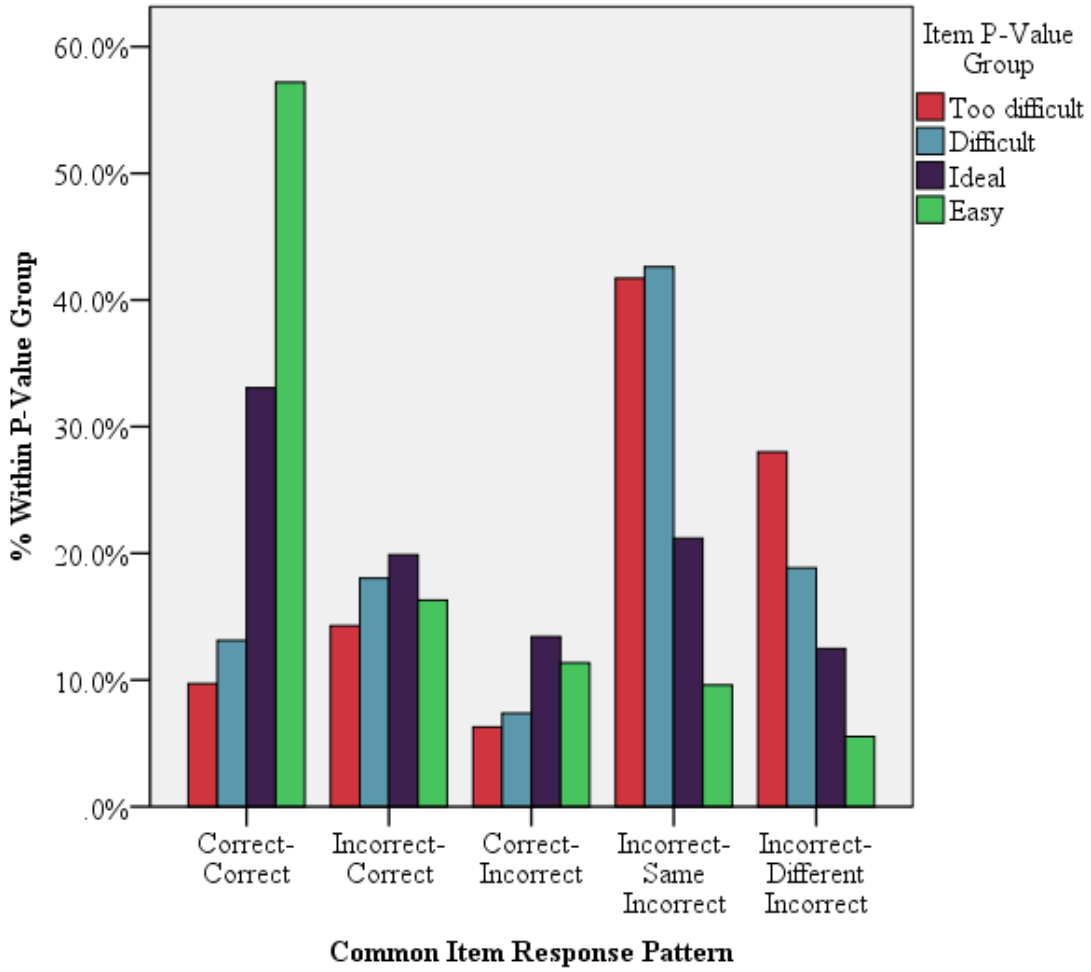

Figure 4. Common item response patterns by item difficulty group.

Table XIII

Response Patterns and Times With Mean Item Difficulty Indices (N = 30,209)

					Response time (seconds)					
	Frequency		Item p-value		Initial attempt		Repeat attempt		Change[a]	
Item type and response pattern	*n*	%	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Common item (*n* = 6,005)										
Correct–Correct	2,144	35.7%	.72	.13	56.19	47.12	55.47	50.89	-0.71	51.35
Incorrect–Correct	1,143	19.0%	.66	.15	76.27	62.72	69.10	60.09	-7.17	72.99
Correct–Incorrect	757	12.6%	.67	.14	70.20	51.02	80.33	59.77	10.13	64.31
Incorrect–Same Incorrect	1,241	20.7%	.58	.17	66.25	48.42	66.10	49.71	-0.15	55.54
Incorrect–Different Incorrect	720	12.0%	.59	.17	76.33	56.16	78.79	57.68	2.46	62.96
Unique item (*n* = 24,204)										
Correct	13,657	56.4%	.76	.17	51.61	44.51	59.06	53.24	—	—
Incorrect	10,547	43.6%	.60	.21	67.75	55.64	74.83	59.35	—	—

[a]Change in response time was calculated for each individual item. Thus, these values are blank for unique items.

Table XIV displays response pattern and time results by item p-value group. On both common and unique items within each p-value group, repeat examinees typically took less time to respond correctly than incorrectly across exam attempts. However, the converse was true on the too difficult items, those with p-values less than .30 among all first-time examinees including non-repeaters who passed. On these items, correct responses took more time while incorrect responses took less. On such challenging items, longer response times might then indicate ultimately successful reasoning processes, whereas shorter response times might be compatible with unlucky blind guesses.

Results for common items as a whole had revealed correct–correct as the most frequent response pattern. This general trend did not hold across the full range of item difficulty. The correct–correct response pattern remained the most frequent only among ideal and easy item groups, suggesting that such common items lent better to persistent knowledge. This was not the case among the difficult and too difficult common item groups, so these items may have been more susceptible to memory disadvantages.

Previous results for all common items had revealed comparable times for correct–correct responses across exam attempts. This finding did not apply across all item difficulty groups. Mean response times were comparable only on ideal and easy common items, again suggesting that ideal and easy common items lent better to persistent knowledge. However, mean response times were more discrepant among the difficult and too difficult common items. On the difficult items, participants spent more time reselecting the correct answer. Perhaps the difficulty of these items hindered the reasoning process. On the too difficult items, participants spent slightly less time reselecting the correct answer. Perhaps the extreme difficulty of such items occasionally motivated post-initial exam attempt research to confirm initial answers.

Table XIV

Response Patterns and Times by Item Difficulty Group

	Frequency		Response time (seconds)					
			Initial attempt		Repeat attempt		Change[a]	
Item type and response pattern	*n*	%	*M*	*SD*	*M*	*SD*	*M*	*SD*
Too difficult (p-value < .30)								
Common item (*n* = 175)								
Correct–Correct	17	9.7%	51.00	43.46	46.88	40.20	-4.12	29.50
Incorrect–Correct	25	14.3%	56.80	39.94	53.08	35.44	-3.72	42.79
Correct–Incorrect	11	6.3%	59.55	29.03	62.82	57.05	3.27	38.64
Incorrect–Same Incorrect	73	41.7%	49.74	38.06	51.25	42.92	1.51	48.43
Incorrect–Different Incorrect	49	28.0%	44.78	30.15	65.88	49.50	21.10	39.16
Unique item (*n* = 929)								
Correct	115	12.4%	70.22	55.23	69.42	52.23	—	—
Incorrect	814	87.6%	57.33	44.82	66.89	52.70	—	—
Difficult (.30 ≤ p-value < .40)								
Common item (*n* = 244)								
Correct–Correct	32	13.1%	63.47	37.78	76.16	54.11	12.69	50.83
Incorrect–Correct	44	18.0%	89.93	62.28	100.23	73.21	10.30	72.11
Correct–Incorrect	18	7.4%	102.72	40.49	94.89	69.39	-7.83	55.91
Incorrect–Same Incorrect	104	42.6%	66.79	48.12	65.75	46.14	-1.04	53.67
Incorrect–Different Incorrect	46	18.9%	64.98	40.38	63.28	45.99	-1.70	58.97
Unique item (*n* = 1,558)								
Correct	417	26.8%	60.09	46.42	77.60	57.24	—	—
Incorrect	1,141	73.2%	64.05	51.14	81.59	63.12	—	—

Table XIV (*continued*)

	Frequency		Response time (seconds)					
			Initial attempt		Repeat attempt		Change[a]	
Item type and response pattern	*n*	%	*M*	*SD*	*M*	*SD*	*M*	*SD*
Ideal (.40 ≤ p-value ≤ .80)								
Common item (*n* = 4,556)								
Correct–Correct	1,506	33.1%	60.60	49.93	59.12	53.69	-1.48	54.75
Incorrect–Correct	906	19.9%	77.56	63.25	70.75	61.69	-6.81	73.54
Correct–Incorrect	611	13.4%	71.56	51.74	80.73	60.44	9.17	65.67
Incorrect–Same Incorrect	965	21.2%	68.12	49.41	67.66	50.90	-0.46	56.79
Incorrect–Different Incorrect	568	12.5%	79.60	58.33	81.60	59.23	2.00	65.71
Unique item (*n* = 13,124)								
Correct	6,546	49.9%	61.27	49.57	67.34	57.55	—	—
Incorrect	6,578	50.1%	69.30	55.24	74.36	57.70	—	—
Easy (p-value > .80)								
Common item (*n* = 1,030)								
Correct–Correct	589	57.2%	44.66	37.39	45.27	41.22	0.60	42.00
Incorrect–Correct	168	16.3%	68.61	61.77	54.41	44.83	-14.20	73.39
Correct–Incorrect	117	11.4%	59.09	47.61	77.65	54.90	18.56	59.64
Incorrect–Same Incorrect	99	9.6%	59.58	43.18	62.26	44.59	2.69	50.34
Incorrect–Different Incorrect	57	5.5%	79.98	53.43	74.39	54.37	-5.60	51.80
Unique item (*n* = 8,593)								
Correct	6,579	76.6%	43.00	37.55	47.08	44.22	—	—
Incorrect	2,014	23.4%	69.96	62.41	75.37	64.93	—	—

Note. *N* response patterns = 30,209; *n* common item patterns = 6,005; *n* unique item patterns = 24,204.

[a]Change in response time was calculated for each individual item. Thus, these values are blank for unique items.

With the incorrect–same incorrect pattern, the results across all common items had revealed similar mean response times across exam attempts. This general observation held for each p-value group. Regardless of item difficulty, participants may have engaged in similar faulty reasoning processes when selecting and reselecting a distractor. Previous results for common items as a whole had also implied increased certainty in reselected distractors versus distractors that participants rationalized only once. This similarly applied to each p-value group. Mean times for incorrect–same incorrect responses were usually shorter than times for other incorrect responses on common and unique items alike.

Item difficulty and false knowledge building may have been related. On easy common items, the incorrect–same incorrect pattern was the least frequent response pattern. On ideal common items, this pattern was the second most frequent. On difficult and too difficult common items, it was the most frequent. Moreover, on ideal, difficult, and too difficult items, participants provided more incorrect–same incorrect responses than they did remediated responses. This was not the case among easy items. Therefore, pattern and time results suggest that increased item difficulty may have better fostered false knowledge building.

Rates of incorrect–correct responses were fairly comparable among item groups, with a 6% range between the highest rate (ideal items) and the lowest rate (too difficult items). However, the corresponding response times suggest some dissimilarity in how repeat examinees demonstrated remediated knowledge across the range of item difficulty. On easy, ideal, and too difficult common items, participants generally decreased response times. In contrast, on difficult common items, remediating responses took a longer time. Therefore, participants may have experienced more of a memory advantage on easy, ideal, and too difficult common items. This does not eliminate the possibility of a memory advantage on difficult common items, but

participants may have also expended more effort on reasoning through the answer on these items.

Participants appeared especially prepared to re-challenge easy common items. Among the easy item group, incorrect–correct was the second most frequent response pattern. Additionally, the easy item group garnered the largest average decrease in response time for remediated responses (M = -14.20 s). The easiness of these common items might have then lent to a memory advantage.

Though the too difficult group saw the lowest rate of remediated responses along with only a small average decrease in time, participants may have sometimes experienced a unique advantage on these items. For the easy, ideal, and difficult item groups, the mean response time for a remediated response was longer than mean times for correct responses on unique items. In contrast, for too difficult items, mean time for a remediated response was lower than the mean times for correct responses on unique items. Again, the extreme difficulty of some items may have made them easier to memorize and research after initial exam attempts.

On difficult and too difficult common items, the rates of correct–incorrect patterns were 6% and 7%, respectively. In comparison, the rates of correct–incorrect patterns on easy and ideal common items were 11% and 13%, respectively. On ideal and especially on easy items, repeat examinees spent considerably more time changing from the correct answer to a distractor. Again, perhaps higher item difficulty occasionally prompted research of an initial correct response or lucky guess. At the same time, item easiness may have more frequently given way to overthinking and subsequent false recall.

Several times I have remarked that item difficulty may have lent to a specific memory advantage, motivation to research and confirm an initially correct answer. Nevertheless, repeat

examinees were largely unable to address knowledge gaps on difficult and too difficult items. Again, remediated responses occurred at a lower rate for these items compared to easy or ideal items. Though difficult items garnered a lower rate of correct–incorrect responses compared to easy or ideal items, participants still appeared to grapple with these items. Unlike on easy or ideal items, such responses on difficult items were accompanied with an average decrease in time, conveying a degree of certainty in these unnecessary corrections. Lastly, for incorrect–different patterns, mean response times were fairly comparable across attempts on difficult items and considerably increased on too difficult items. Such time results are compatible with unsuccessful remediation attempts rather than, for instance, blind guessing on repeat attempts. These results altogether suggest that more extreme item difficulty did not lend well to remediating lack of knowledge. Instead, any indication of a relationship between difficulty and a studying advantage was evident only on occasion and only on items that participants were able to initially answer correctly.

c. By item discrimination index

Next, I present response patterns and times on items grouped by the levels of discrimination I defined for Question 2(a). As a reminder, most items had acceptable point-biserial correlations, though several items had negative point-biserial correlations. The high discrimination group was comprised of items with a point-biserial correlation of .15 or above and items with a p-value of .90 or above and a positive point-biserial correlation. The low discrimination group consists of items with positive point-biserial correlations less than .15. Figure 5 shows somewhat similar rates of each common item response pattern across all item discrimination groups. On average, item discrimination was similar across observed item response patterns (see Table XV).

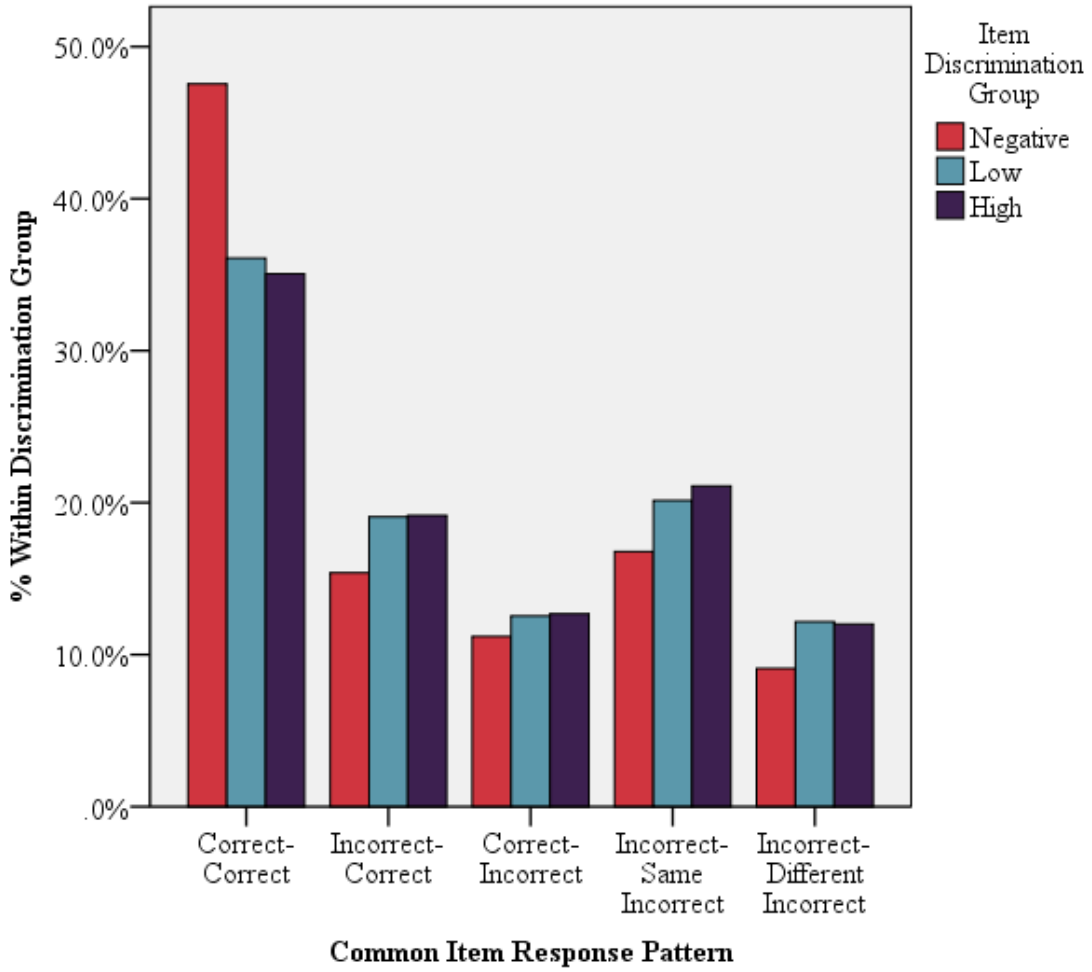

Figure 5. Common item response patterns by item discrimination group.

Table XV

Response Patterns and Times With Mean Item Discrimination Indices (N = 30,209)

					Response time (seconds)					
	Frequency		Item r_{pbis}		Initial attempt		Repeat attempt		Change[a]	
Item type and response pattern	*n*	%	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Common item (*n* = 6,005)										
Correct–Correct	2,144	35.7%	.16	.09	56.19	47.12	55.47	50.89	-0.71	51.35
Incorrect–Correct	1,143	19.0%	.17	.09	76.27	62.72	69.10	60.09	-7.17	72.99
Correct–Incorrect	757	12.6%	.17	.09	70.20	51.02	80.33	59.77	10.13	64.31
Incorrect–Same Incorrect	1,241	20.7%	.17	.09	66.25	48.42	66.10	49.71	-0.15	55.54
Incorrect–Different Incorrect	720	12.0%	.18	.09	76.33	56.16	78.79	57.68	2.46	62.96
Unique item (*n* = 24,204)										
Correct	13,657	56.4%	.17	.11	51.61	44.51	59.06	53.24	—	—
Incorrect	10,547	43.6%	.17	.11	67.75	55.64	74.83	59.35	—	—

[a]Change in response time was calculated for each individual item. Thus, these values are blank for unique items.

However, evaluating response patterns and times by item discrimination group indicates that repeat examinees differentially approached items of varying discrimination (see Table XVI). On both the common and unique items that positively discriminated among all first-time certification examinees, participants generally took less time to respond correctly and more time to respond incorrectly across exam attempts. In contrast was the negative discrimination item group. For this item group, the shortest times corresponded to incorrect–different incorrect responses. Also, the response time discrepancy between correct and incorrect responses on unique items was more pronounced on positively discriminating items than on negatively discriminating items. Therefore, when challenging negatively discriminating items, repeat examinees may have felt fairly certain of the distractors they selected on either attempt or may have unsuccessfully blindly guessed.

Previous results for common items as a whole showed that participants most frequently reselected the correct answer using about the same amount of time across attempts. Both findings held true for each item discrimination group. Magnitude of mean time differences ranged from 0.21 to 2.60 seconds. Within each level of item discrimination, repeat examinees may have then employed similar reasoning processes when demonstrating persistent knowledge.

Comparing item groups, the negative discrimination group showed the highest rate of persistent knowledge. This is not unexpected. A negative point-biserial correlation indicates that examinees who earned lower overall scores when first taking the exam, such as the repeat examinees, were more likely than high scoring examinees to answer the item correctly. The rates for correct–correct responses were similar between the low and high discrimination groups.

Table XVI

Response Patterns and Times by Item Discrimination Group

	Frequency		Response time (seconds)					
			Initial attempt		Repeat attempt		Change[a]	
Item type and response pattern	*n*	%	*M*	*SD*	*M*	*SD*	*M*	*SD*
Negative ($r_{pbis} < 0$)								
Common item (*n* = 143)								
Correct–Correct	68	47.6%	49.71	32.79	47.10	35.20	-2.60	35.13
Incorrect–Correct	22	15.4%	66.95	44.33	59.23	45.24	-7.73	65.02
Correct–Incorrect	16	11.2%	57.69	25.00	58.13	55.12	0.44	42.78
Incorrect–Same Incorrect	24	16.8%	55.04	47.90	65.63	65.98	10.58	73.66
Incorrect–Different Incorrect	13	9.1%	44.08	30.89	48.54	27.19	4.46	27.20
Unique item (*n* = 845)								
Correct	465	55.0%	54.48	38.89	67.39	50.81	—	—
Incorrect	380	45.0%	61.68	43.77	69.65	51.69	—	—
Low ($0 \leq r_{pbis} < .15$)[b]								
Common item (*n* = 2,040)								
Correct–Correct	736	36.1%	56.62	46.28	54.40	43.80	-2.22	44.55
Incorrect–Correct	389	19.1%	78.10	66.34	72.80	62.67	-5.30	75.42
Correct–Incorrect	256	12.5%	66.75	43.96	76.82	51.19	10.07	52.66
Incorrect–Same Incorrect	411	20.1%	65.45	48.55	66.84	48.27	1.39	56.20
Incorrect–Different Incorrect	248	12.2%	78.71	58.70	82.95	64.87	4.24	61.39
Unique item (*n* = 7,573)								
Correct	3,948	52.1%	55.32	46.55	65.88	58.43	—	—
Incorrect	3,625	47.9%	69.35	59.49	78.38	61.14	—	—

Table XVI (*continued*)

	Frequency		Response time (seconds)					
			Initial attempt		Repeat attempt		Change[a]	
Item type and response pattern	*n*	%	*M*	*SD*	*M*	*SD*	*M*	*SD*
			High ($r_{pbis} \geq .15$)					
Common item (*n* = 3,822)								
Correct–Correct	1,340	35.1%	56.28	48.18	56.49	55.00	0.21	55.37
Incorrect–Correct	732	19.2%	75.57	61.22	67.43	59.03	-8.14	71.96
Correct–Incorrect	485	12.7%	72.43	54.89	82.92	63.84	10.49	70.27
Incorrect–Same Incorrect	806	21.1%	66.98	48.38	65.74	49.95	-1.25	54.60
Incorrect–Different Incorrect	459	12.0%	75.96	55.09	77.40	53.83	1.44	64.56
Unique item (*n* = 15,786)								
Correct	9,244	58.6%	49.74	43.73	56.03	50.91	—	—
Incorrect	6,542	41.4%	67.08	53.69	73.40	58.85	—	—

Note. *N* response patterns = 30,209; *n* common item patterns = 6,005; *n* unique item patterns = 24,204.

[a]Change in response time was calculated for each individual item. Thus, these values are blank for unique items. [b]Excludes items with p-values greater than .90. Items with p-values greater than .90 and any positive point-biserial correlation were classified as High.

Results for common items as a whole had revealed incorrect–same incorrect as the second most frequent common item response pattern. This general trend applied to each discrimination group. Previous results had also indicated comparable response times for incorrect–same incorrect responses across exam attempts. This held only for positively discriminating items. On negatively discriminating items, however, participants spent an average of 10.58 seconds longer to reselect a distractor than they did when initially select it. Persistent false knowledge may have then carried different burdens in relation to item discrimination. On positively discriminating items, repeat examinees may have similarly applied faulty reasoning across exam attempts. On negatively discriminating items, repeat examinees spent more time and therein more mental energy toward ultimately settling on the same distractor.

Though participants somewhat less frequently demonstrated persistent knowledge on positively discriminating items than on negatively discriminating items, they demonstrated remediated knowledge on the positively discriminating items at slightly higher rates. Rates of remediated responses were similar between the low and high discrimination groups. Across all discrimination groups, incorrect–correct responses were accompanied with a mean decrease in time, suggesting that participants experienced a memory advantage on common items across the range of item discrimination.

As item discrimination improved, the rate of correct–incorrect response patterns only slightly increased. Again, this is not unexpected as lower scoring examinees are more likely to answer negatively or weakly discriminating items correctly. Ultimately, all three item groups had similar rates of correct–incorrect patterns.

However, mean differences in time for correct–incorrect responses varied across item groups. For both the low and high discrimination groups, participants spent an average of about 10 seconds longer to change from the correct answer to a distractor. This additional response time suggests increased deliberation and overthinking. For the negative group, however, participants used roughly the same amount of time to change from the correct answer to a distractor. Participants did not appear to ponder their initial correct responses in the same manner they may have when re-challenging positively discriminating items. On 12 out of the 16 correct–incorrect responses on negatively discriminating items, participants had selected a distractor with a p-value of .10 or greater and a positive point-biserial correlation. Specifically, they had instead changed from the keyed answer to a distractor that attracted even some high scoring first-time examinees, illustrating one pitfall with reusing negatively discriminating items even if SMEs deem those items valid with respect to content.

Rate of incorrect–different incorrect responses was similar between the low and high discrimination groups and slightly lower for the negative group. On positively discriminating items, mean times for incorrect–different incorrect responses were typically longer than times for other response types on common or unique items alike. This suggests a considerable amount of time and effort to work through a persistent lack of knowledge. On negatively discriminating common items, mean times were shorter for incorrect–different incorrect responses than for most of the other response types on common or unique items. Again, item flaws such as ambiguity may have prompted repeat examinees to confidently provide incorrect responses or quickly provide blind guesses.

d. By content area/cognitive complexity subtest

Next, I present response patterns and times on items grouped by content area and cognitive complexity level. As a reminder, the subtests of items were basic biomedical science recall, basic biomedical science application, clinical science recall, and clinical science application. Frequency of common item response pattern was significantly associated with subtest, χ^2 (12, N = 6,005) = 66.21, $p < .001$. This is unsurprising given that the participants had scored differently across subtests on both exam attempts. Figure 6 displays the rates of common item response patterns by each subtest. The most frequent response pattern for each subtest was correct–correct, indicating that repeat examinees demonstrated persistent knowledge across the range of content areas and skills represented on the certification exam. Moreover, each subtest saw considerable rates of the remaining response patterns. That is, repeat examinees appeared to experience both memory advantages and disadvantages on each subtest.

Table XVII displays response pattern and time results by content area/cognitive complexity subtest. On both common and unique items within each subtest, repeat examinees generally took less time to respond correctly than incorrectly on either exam attempt. Repeat examinees also generally took less time to respond to a recall item than to an application item. Application items tended to be longer than recall items because they contained more information for examinees to interpret in order to respond, which might help explain this time difference.

Similar to previous results for common items as a whole, participants most frequently reselected the correct answer across exam attempts on each of the four subtests. Mean response times for this pattern were somewhat comparable, with magnitude of mean differences ranging from 0.55 to 2.38 seconds. Therefore, repeat examinees most frequently demonstrated persistent knowledge across the range of content and skills on the exam. The comparable response times

suggest that they may have engaged in somewhat similar reasoning processes when demonstrating this persistent knowledge on each subtest. Participants demonstrated persistent knowledge at slightly higher rates on the two application subtests than they did on the two recall subtests. This bolsters the finding from Question 2(b) indicating that participants tended to demonstrate more knowledge on application items.

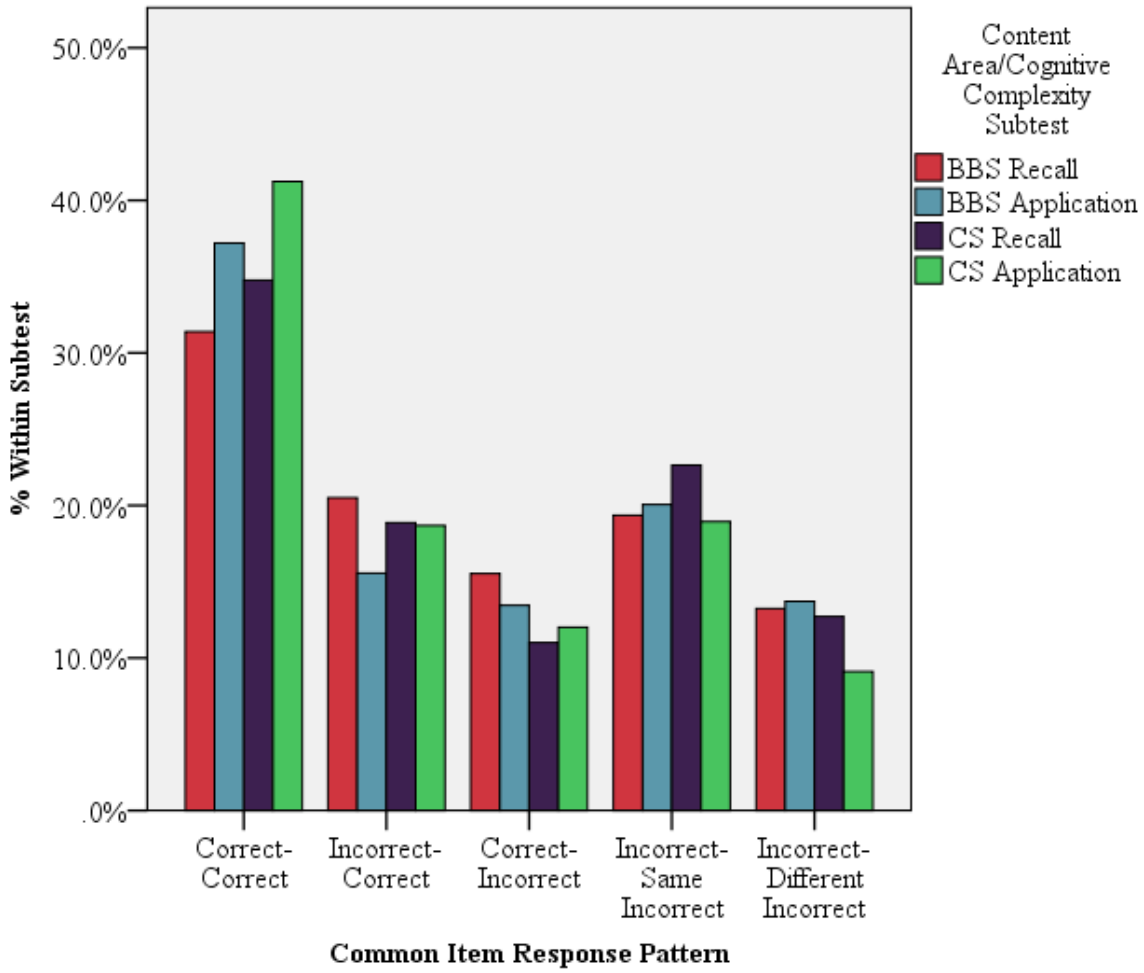

Figure 6. Common item response patterns by content area/cognitive complexity subtest. BBS = basic biomedical science; CS = clinical science.

Table XVII

Response Patterns and Times by Content Area/Cognitive Complexity Subtest

	Frequency		Response time (seconds)					
			Initial attempt		Repeat attempt		Change[a]	
Item type and response pattern	*n*	%	*M*	*SD*	*M*	*SD*	*M*	*SD*
Basic Biomedical Science Recall								
Common item (*n* = 1,571)								
Correct–Correct	493	31.4%	47.47	37.73	48.01	47.32	0.55	52.14
Incorrect–Correct	322	20.5%	62.99	43.91	60.69	51.88	-2.30	62.67
Correct–Incorrect	244	15.5%	60.83	44.08	71.98	56.14	11.15	61.43
Incorrect–Same Incorrect	304	19.4%	63.29	50.00	57.82	42.99	-5.47	59.76
Incorrect–Different Incorrect	208	13.2%	65.51	43.11	70.29	46.95	4.78	59.13
Unique item (*n* = 7,104)								
Correct	3,748	52.8%	43.83	38.09	47.73	42.60	—	—
Incorrect	3,356	47.2%	59.74	45.64	67.95	53.25	—	—
Basic Biomedical Science Application								
Common item (*n* = 379)								
Correct–Correct	141	37.2%	69.11	51.16	71.49	59.47	2.38	56.85
Incorrect–Correct	59	15.6%	108.86	95.96	91.56	72.78	-17.31	98.13
Correct–Incorrect	51	13.5%	95.53	69.77	95.69	60.63	0.16	76.03
Incorrect–Same Incorrect	76	20.1%	82.32	52.89	74.57	44.26	-7.75	50.76
Incorrect–Different Incorrect	52	13.7%	117.77	63.26	100.33	58.68	-17.44	71.98
Unique item (*n* = 1,906)								
Correct	1,046	54.9%	82.68	57.47	75.99	60.13	—	—
Incorrect	860	45.1%	97.02	77.63	90.38	63.66	—	—

Table XVII (*continued*)

			Response time (seconds)					
	Frequency		Initial attempt		Repeat attempt		Change[a]	
Item type and response pattern	*n*	%	*M*	*SD*	*M*	*SD*	*M*	*SD*
Clinical Science Recall								
Common item (*n* = 2,508)								
Correct–Correct	872	34.8%	42.55	32.00	41.74	37.85	-0.81	42.27
Incorrect–Correct	473	18.9%	61.31	47.33	52.16	42.92	-9.15	56.08
Correct–Incorrect	276	11.0%	58.08	42.27	69.76	52.86	11.68	58.12
Incorrect–Same Incorrect	568	22.6%	54.53	38.48	57.05	45.12	2.53	49.27
Incorrect–Different Incorrect	319	12.7%	62.03	43.74	63.60	47.15	1.57	54.31
Unique item (*n* = 9,038)								
Correct	5,006	55.4%	40.66	34.02	46.58	40.54	—	—
Incorrect	4,032	44.6%	56.36	42.65	61.10	47.22	—	—
Clinical Science Application								
Common item (*n* = 1,547)								
Correct–Correct	638	41.2%	78.71	59.47	76.47	58.71	-2.24	60.10
Incorrect–Correct	289	18.7%	108.88	77.93	101.61	74.27	-7.27	97.98
Correct–Incorrect	186	12.0%	93.52	55.28	102.77	66.95	9.25	72.98
Incorrect–Same Incorrect	293	18.9%	87.87	54.36	90.04	57.39	2.17	62.86
Incorrect–Different Incorrect	141	9.1%	109.36	72.58	117.75	72.06	8.39	80.06
Unique item (*n* = 6,156)								
Correct	3,857	62.7%	68.52	51.93	78.59	64.43	—	—
Incorrect	2,299	37.3%	94.69	71.28	98.16	71.84	—	—

Note. *N* response patterns = 30,209; *n* common item patterns = 6,005; *n* unique item patterns = 24,204.

[a]Change in response time was calculated for each individual item. Thus, these values are blank for unique items.

On each subtest, incorrect–correct responses were accompanied with a mean decrease in time, signifying knowledge remediation and potential memory advantages across the content areas and skills on the exam. Comparing subtests, participants remediated responses on basic biomedical science recall items at a slightly higher rate compared to the other subtests. However, the corresponding mean decrease in time was small (M = -2.30 s), perhaps reflecting some level of effort to summon any knowledge accrued when preparing to retest. Rates of remediated responses between the two clinical science subtests were comparable, as were corresponding mean decreases in time. The basic biomedical science application subtest saw the lowest rate of remediated responses, yet the corresponding mean decrease in time was the largest among all subtests (M = -17.31 s). As the basic biomedical application subtest was the least represented area on the exam, it may have well been the least representative of medical practice. In turn, repeat examinees may have been comparatively unfamiliar with the underlying concepts of these items, helping to explain the slightly lower rate of remediated responses on this subtest. Yet when participants did remediate their responses, they did so in substantially shorter time. Greater unfamiliarity with these items may have made them more memorable, thereby lending to a possible studying advantage.

The rates of correct–incorrect patterns were slightly higher on the basic biomedical science subtests than on the clinical science subtests, suggesting that basic biomedical science content may have lent somewhat better to a memory disadvantage such as false recall. Earlier results for common items as a whole showed that participants generally spent somewhat more time changing from the correct answer to a distractor. These general trends held for the basic biomedical science recall, clinical science recall, and clinical science application subtests, implying increased deliberation and overthinking. However, on the basic biomedical science

application subtest, response times were comparable across exam attempts. In this instance, unfamiliarity with these items may have been a detriment to retest performance.

The clinical science application subtest garnered the smallest rate of incorrect–different incorrect responses among all subtests. Rates of incorrect–different incorrect responses were comparable between the basic biomedical science recall, basic biomedical science application, and clinical science recall subtests. However, response time results varied. The basic biomedical science application subtest yielded the only mean decrease in response time, and the time decrease was considerable. This time decrease might suggest misguided remediation efforts on this item type. Alternatively, they may have guessed on both exam attempts after spending a lot of time trying to determine the answer, more quickly guessing the second time around. Again, unfamiliarity with the concepts that these items tested might help explain why this subtest did not lend as well to remediating lack of knowledge.

e. Among passing examinees with score gains beyond measurement error

For Question 1(c), I had identified six examinees who had passed their repeat attempts with score gains beyond measurement error. All examinees in this subgroup had a one-year interval between exam attempts. To more closely investigate whether a memory advantage from prior item exposure may have related to their exceeding score gains, I evaluated their common item response patterns and mean response times individually and as a group.

Figure 7 displays rates of each common item response pattern across these examinees as a group, as well as the rates for failing examinees and for those passing with score gains within measurement error. Compared to failing examinees, both groups of passing examinees had higher rates of correct–correct and incorrect–correct patterns. Between both groups of passing examinees, those who passed with score gains beyond measurement error demonstrated

persistent knowledge and remediated knowledge at only slightly higher rates than those who passed with more modest score gains. Both groups of passing examinees also demonstrated persistent false knowledge at similar rates, though these rates were considerably lower than that for failing examinees. Examinees passing with exceeding score gains had the lowest rates of correct–incorrect and incorrect–different incorrect patterns than examinees who either failed or passed with score gains within measurement error. Therefore, they demonstrated potential false recall and erroneous reasoning at lower rates.

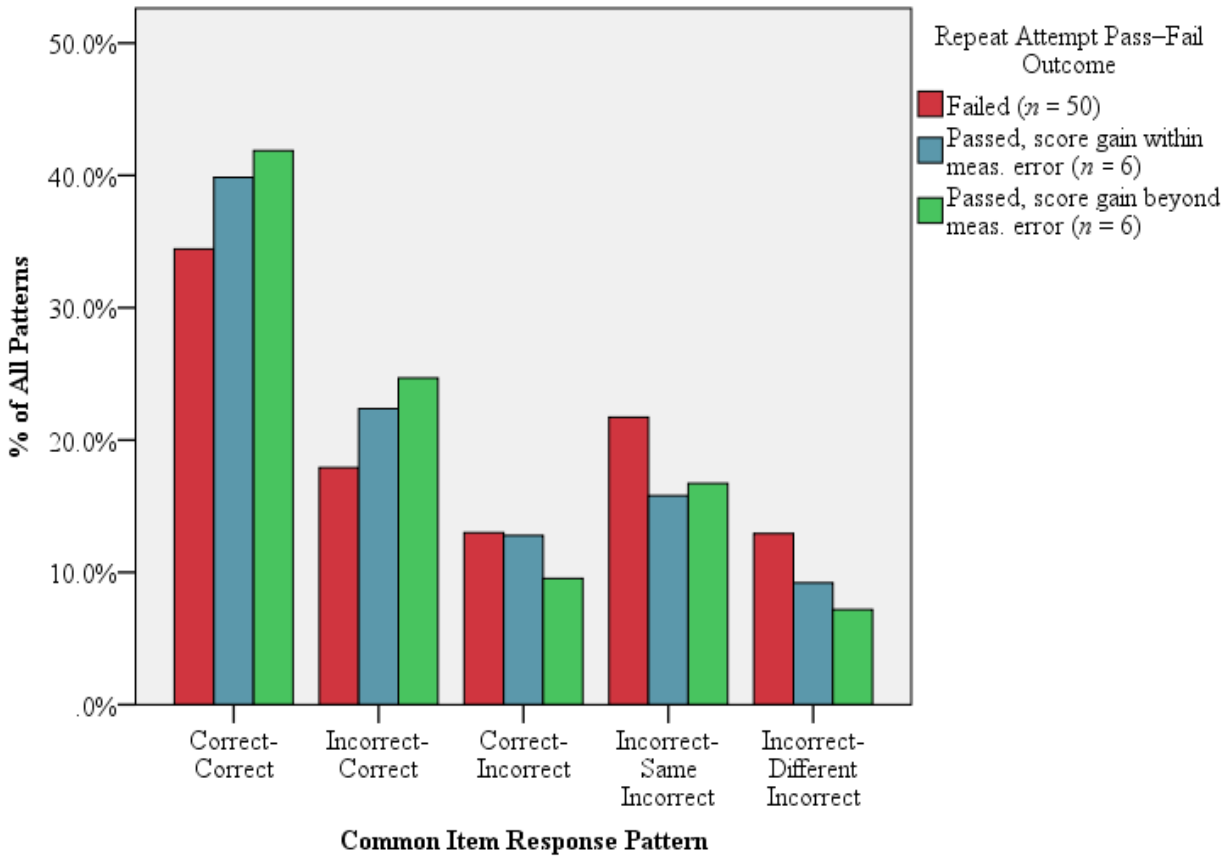

Figure 7. Common item response patterns by retest pass–fail group.

Table XVIII displays group-level response patterns and mean response times for the examinees who passed with score gains beyond measurement error, both across all common items and by content area/cognitive complexity subtest. Results for this subgroup of passing participants suggest high rates of persistent knowledge. Similar to previous results for all study participants, the most frequent common item response pattern among the subgroup was correct–correct. This applied to each individual in the subgroup, with percentage of correct–correct responses ranging from 35% to 49% and averaging about 42% of all common item response patterns. In addition, correct–correct was the most frequent response pattern on each content area/cognitive complexity subtest (see Figure 8). Therefore, participants who passed with more dramatic score gains demonstrated persistent knowledge across the range of content and skills represented on the certification exam, thereby lending support to their pass outcomes.

Previous results for all study participants had revealed incorrect–same incorrect as the second most frequent response pattern after correct–correct. In contrast, for five out of the six examinees who passed with a score gain beyond measurement error, incorrect–correct was the second most frequent response pattern after correct–correct. For the individual who did not adhere to this trend, the rate of incorrect–correct patterns was 3% less than the rate of incorrect–same incorrect patterns, his or her second most frequent response pattern. Individual rates of reselecting distractors ranged from 10% to 22% of all common item response patterns, whereas individual rates of remediated responses ranged from 19% to 28%. Though examinees in the subgroup demonstrated persistent false knowledge across common items as a whole, they generally did so less often than they remediated responses.

Table XVIII

Response Patterns and Times for Passing Examinees With Score Gains Beyond Measurement Error (N = 2,750)

			Response time (seconds)					
	Frequency		Initial attempt		Repeat attempt		Change[a]	
Item type and response pattern	*n*	%	*M*	*SD*	*M*	*SD*	*M*	*SD*
All scored items								
Common item (*n* = 640)								
Correct–Correct	268	41.9%	60.29	48.43	66.96	56.13	6.66	51.17
Incorrect–Correct	158	24.7%	85.71	73.11	64.70	43.48	-21.01	76.85
Correct–Incorrect	61	9.5%	92.02	61.54	110.89	71.47	18.87	64.64
Incorrect–Same Incorrect	107	16.7%	72.14	50.10	76.58	44.93	4.44	55.94
Incorrect–Different Incorrect	46	7.2%	79.83	49.94	88.72	51.77	8.89	66.39
Unique item (*n* = 2,110)								
Correct	1,339	63.5%	59.55	52.41	61.08	51.36	—	—
Incorrect	771	36.5%	83.71	66.48	92.09	58.32	—	—
Basic Biomedical Science Recall								
Common item (*n* = 196)								
Correct–Correct	84	42.9%	49.32	40.67	54.67	47.12	5.35	51.73
Incorrect–Correct	55	28.1%	70.76	42.82	58.93	39.98	-11.84	53.35
Correct–Incorrect	22	11.2%	83.27	50.58	104.50	58.63	21.23	70.85
Incorrect–Same Incorrect	21	10.7%	64.00	40.45	76.19	48.05	12.19	53.23
Incorrect–Different Incorrect	14	7.1%	69.57	43.73	77.71	49.26	8.14	63.03
Unique item (*n* = 547)								
Correct	337	61.6%	47.16	40.31	46.90	42.24	—	—
Incorrect	210	38.4%	69.37	49.56	101.72	65.12	—	—

Table XVIII (*continued*)

	Frequency		Response time (seconds)					
			Initial attempt		Repeat attempt		Change[a]	
Item type and response pattern	*n*	%	*M*	*SD*	*M*	*SD*	*M*	*SD*
Basic Biomedical Science Application								
Common item (*n* = 45)								
Correct–Correct	24	53.3%	69.25	41.71	70.46	49.06	1.21	45.74
Incorrect–Correct	6	13.3%	205.00	218.69	124.50	100.60	-80.50	264.26
Correct–Incorrect	5	11.1%	127.60	83.98	143.00	97.49	15.40	92.99
Incorrect–Same Incorrect	8	17.8%	108.13	62.87	78.63	41.94	-29.50	65.06
Incorrect–Different Incorrect	2	4.4%	149.50	78.49	128.00	8.49	-21.50	86.97
Unique item (*n* = 176)								
Correct	116	65.9%	89.9	56.61	72.58	51.06	—	—
Incorrect	60	34.1%	102.21	72.57	91.66	45.21	—	—
Clinical Science Recall								
Common item (*n* = 230)								
Correct–Correct	82	35.7%	42.40	30.80	52.16	45.08	9.76	43.39
Incorrect–Correct	62	27.0%	70.40	41.48	56.45	37.87	-13.95	48.11
Correct–Incorrect	20	8.7%	61.75	41.77	87.85	77.73	26.10	60.59
Incorrect–Same Incorrect	47	20.4%	59.04	41.74	67.19	48.49	8.15	52.73
Incorrect–Different Incorrect	19	8.3%	75.21	48.55	78.79	51.57	3.58	71.57
Unique item (*n* = 807)								
Correct	501	62.1%	51.66	50.48	48.20	42.63	—	—
Incorrect	306	37.9%	70.24	49.44	80.98	55.22	—	—

Table XVIII (*continued*)

	Frequency		Response time (seconds)					
			Initial attempt		Repeat attempt		Change[a]	
Item type and response pattern	*n*	%	*M*	*SD*	*M*	*SD*	*M*	*SD*
Clinical Science Application								
Common item (*n* = 169)								
Correct–Correct	78	46.2%	88.15	59.56	94.67	67.05	6.51	59.73
Incorrect–Correct	35	20.7%	115.86	86.12	78.14	33.70	-37.71	84.81
Correct–Incorrect	14	8.3%	136.29	67.01	142.36	62.62	6.07	53.65
Incorrect–Same Incorrect	31	18.3%	88.23	57.06	90.55	35.15	2.32	59.52
Incorrect–Different Incorrect	11	6.5%	88.18	51.39	112.73	53.01	24.55	64.71
Unique item (*n* = 580)								
Correct	385	66.4%	75.74	57.78	83.90	58.58	—	—
Incorrect	195	33.6%	115.36	88.50	106.57	60.70	—	—

[a]Change in response time was calculated for each individual item. Thus, these values are blank for unique items.

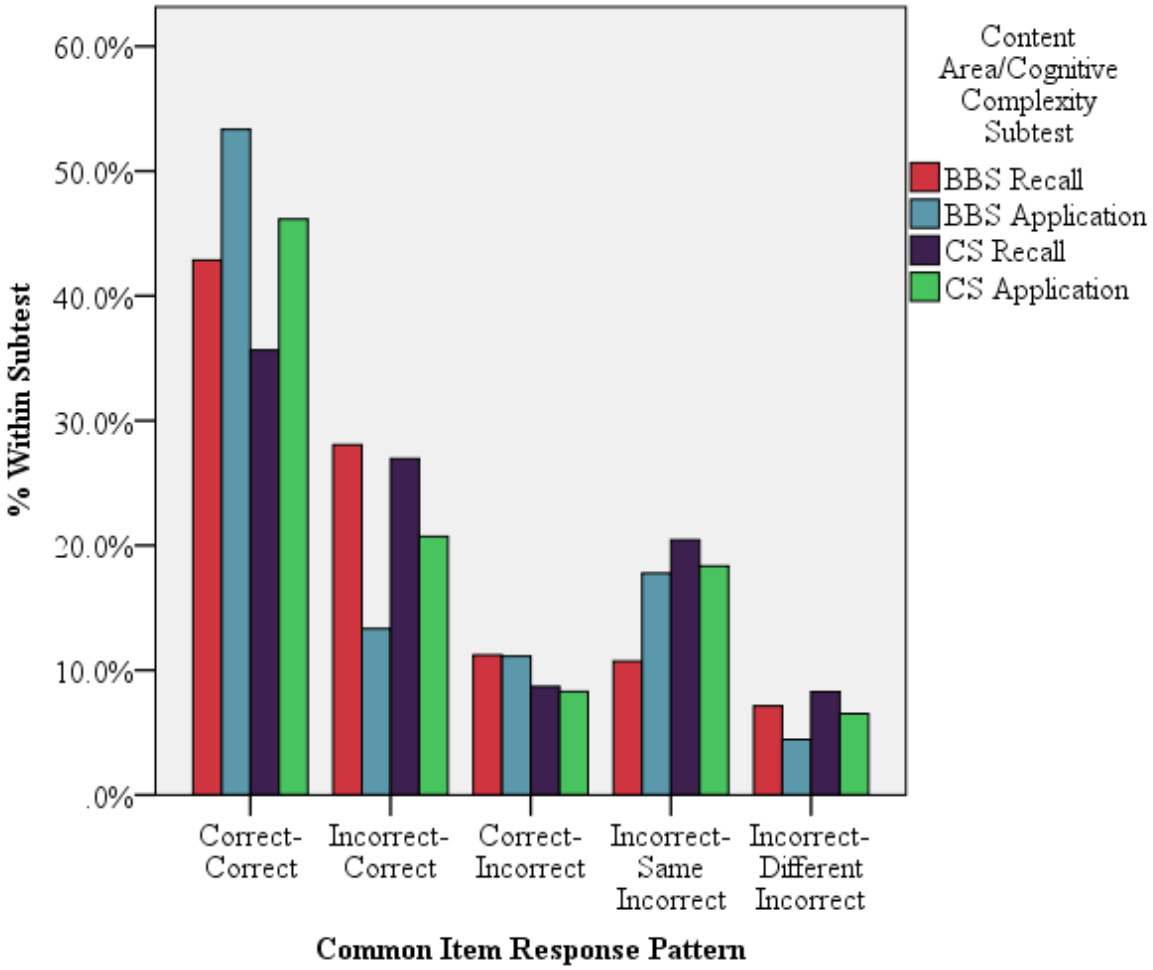

Figure 8. Common item response patterns by content area/cognitive complexity subtest for passing examinees with score gains beyond measurement error. BBS = basic biomedical science; CS = clinical science.

Examinees in the current subgroup remediated responses across all subtests. The basic biomedical science recall subtest garnered the highest rate of remediated responses, as well as the lowest rate of reselected distractors. In comparison, the basic biomedical science application subtest saw the lowest rate of remediated responses. Basic biomedical science recall items may have better lent to knowledge remediation efforts, whereas basic biomedical science application items may have better lent to false knowledge building. Regarding the two clinical science subtests, rates of incorrect–correct and incorrect–same incorrect patterns were comparable. Overall, examinees in this subgroup demonstrated improved performance across all content areas and cognitive complexity levels reflected on the exam.

Results for this subgroup of passing examinees suggest a possible memory advantage due to prior item exposure, though evaluation of unique item results helps put this advantage into perspective. First, examinees tended to decrease response time when they changed from a distractor to the correct response, suggesting an advantage from prior knowledge of the common item. Such time decreases were of larger magnitude compared to those among study participants as a whole. Therefore, examinees in this subgroup were able to recognize the correct answer and remediate initial responses much more quickly. Also, as a group, these examinees more frequently remediated responses on recall subtests than on application subtests. As I previously mentioned, recall items might be easier to memorize and research compared to application items. At the same time, each examinee correctly answered unique items in comparable proportions across the four subtests. This is compatible with examinees having remediated their overall content knowledge, helping to explain how they achieved such large score gains. Together, common and unique item results among this subgroup of passing participants point to remediated content knowledge of the concepts underlying common and unique items alike.

Correct–incorrect response patterns were relatively infrequent, though examinees responded in this manner at similar rates across the four subtests. Individually, the rate of this response pattern ranged from 7% to 13% of all common item response patterns. On initial attempts, the corresponding mean response times were usually longer than mean response times for other initial correct responses. On every subtest, this change was generally accompanied with a mean increase in time. Therefore, in switching from the correct answer to a distractor, examinees may have succumbed to initial doubts regarding their responses.

Examinees in this subgroup exhibited persistent knowledge gaps across the range of content and skills on the certification exam, yet at much smaller rates than they exhibited remediated knowledge gaps. Incorrect–different incorrect responses accounted for a small proportion of common item responses. For four of the six examinees in this subgroup, it was the least frequent response pattern. It was also the least frequent common item response pattern on each subtest. Generally, these response patterns were accompanied with a mean increase in time, suggesting that these examinees usually attempted to reason through these items rather than blindly guess. Notable exceptions occurred on the basic biomedical science application subtest, which garnered some of the largest decreases in response time. Given these time decreases, examinees in this subgroup may have been more likely to give up and unsuccessfully guess on this subtest than on other subtests. Again, potential unfamiliarity with the concepts that these items tested might help explain why this subtest may have better related to unaddressed knowledge gaps, even among those who passed with score gains beyond measurement error.

V. DISCUSSION

The purpose of the current study was to better understand how medical certification repeat examinees demonstrated their initial level of knowledge on first exam attempts and the extent to which they demonstrated sufficiently remediated knowledge on repeat attempts. Because the multiple choice format is the ubiquitous format of written certification exams, this study focused on examinees who repeated a multiple choice certification exam administered by a single medical specialty board. To fully compare how these examinees demonstrated their respective levels of knowledge on each exam attempt, I have contextualized their observed score differences and pass–fail outcomes by analyzing their performances on the overall exam, their performances on different types of items on the exam, and the measurement capabilities of the items repeatedly used to assess their professional competence.

In the previous chapter, I described the results of these different analyses. In this concluding chapter, I integrate and interpret those results for a larger picture of medical certification repeat examinee performance. First, I summarize the results that I had described in Chapter IV. When applicable, I relate my results to the results of previous studies. Next, I discuss the practical implications of these results on medical certification retest policies. After, I evaluate the significance of this research in terms theoretical and practical contributions. To qualify the key findings and ensuing implications from this study, I next describe the limitations of this study. Finally, I offer suggestions for future research to deepen understanding of repeat examinee performances in a high-stakes certification environment.

A. Summary of Research Findings

The previous studies on credentialing repeat examinees have largely explored retest score performances with respect to prior item exposure. Consensus among these studies has indicated

that even though examinees typically improved their scores when repeating a credentialing multiple choice exam, prior item exposure generally did not provide them with an unfair score advantage (e.g., Feinberg et al., 2015; Geving et al., 2015; Raymond et al., 2007, 2009; Wood, 2009). Previous researchers had also found that repeat examinees tended to reproduce initial errors on second exam attempts (Feinberg et al., 2015; O'Neill et al., 2015; Wood, 2009). Through additional examination of the quality of reused multiple choice items and of repeat examinee performances on the total exam and on different types of exam items, the current study has substantiated, qualified, and built upon the conclusions of these earlier studies. In this section, I summarize the key findings for each main research question that guided this study.

1. Research Question 1

   Which repeat examinees' scores were most amenable to change between initial and repeat exam attempts, and to what extent did score differences indicate sufficiently remediated knowledge?

   a. For repeat examinees who initially borderline failed and examinees who initially clearly failed the exam, do overall exam scores differ significantly between initial and repeat exam attempts?
   b. What is the pass rate among repeat examinees?
   c. Among passing repeat examinees, how many score gains are beyond measurement error?

As a group, repeat examinees in the study sample significantly improved their overall exam scores when repeating the certification exam. This result is consistent with previous research on examinees who repeated multiple choice medical credentialing exams (Feinberg et al., 2015; O'Neill et al., 2015; Raymond et al., 2007, 2009; Wood, 2009). At the same time,

most individual retest scores were not significantly different from initial scores, with 95% confidence. The majority of individuals, from both the borderline failing and clearly failing initial performance groups, improved their overall exam scores when retaking the exam. Most score gains were within measurement error, with 95% confidence. A considerable proportion of repeat examinees also experienced score losses, similarly within measurement error.

The similarity of scores across exam attempts is reflected in the retest pass rate of 19%. Despite the high frequency of score gains, most gains were not sufficient for passing. That is, most repeat examinees did not sufficiently remediate their respective level of overall content knowledge. This retest pass rate was considerably lower than pass rates among all certification examinees, a finding that is consistent with results from previous studies on medical credentialing exams (Feinberg et al., 2015; O'Neill et al., 2000; Raymond & Luciw-Dubas 2010; Raymond et al., 2007).

Previous studies suggest that examinees of higher ability are better equipped to learn from their initial exam attempts and sufficiently remediate their knowledge than are examinees of lower ability (Kulik et al., 1984; O'Neill et al., 2000; Wood, 2009). This assertion did not fully hold in the current study. Like in Wood's 2009 study, the repeat examinees who passed in this study had, on average, demonstrated more knowledge on the initial exam attempts compared to those who failed their repeat attempts. Unlike in the 2000 study by O'Neill et al., the passing examinees in this study came from both the borderline failing and clearly failing initial performance groups. With only two of the six borderline examinees passing, not every individual who initially demonstrated more remediable knowledge ended up sufficiently addressing their knowledge gaps upon retesting.

If either of the two examinees who passed their repeat attempts had failed their initial attempts due to measurement error, then the opportunity to retest helped rectify their initial false negative outcomes. Otherwise, they may have simply been more prepared to remediate their comparatively modest knowledge gaps. The four remaining borderline examinees may have been unable to pass due to a number of reasons. Upon retesting, measurement error may have again negatively impacted some of their scores, in which case retesting was unable to resolve their initial misclassifications. For others, their level of knowledge may truly lie just below the passing standard, in which case retesting fortunately did not result in false positive certifications. If some of them felt overly confident after their initial attempts, they may not have made the effort required to reflect on their initial testing experiences (O'Neill et al., 2015). Perhaps seeing they had just missed the pass point on their initial attempts gave them less cause to thoroughly reflect on their content knowledge gaps and devote more time to retest preparation. After all, the persistence of errors on common items was a frequent occurrence among all repeat examinees in this study. I discuss this more in depth when I summarize the findings for Question 3(b).

Half of the repeat examinees who passed did so with a retest score beyond measurement error. With retest scores beyond 95% probability limits, the probability of their retest scores occurring was 5% or less. In the Method chapter, I had hypothesized that such score gains might signify an unfair memory advantage from prior item exposure and thereby potential false positive certifications. Consequently, I more closely examined these individuals' retest performances throughout the study. In doing so, I found evidence to support their sufficiently remediated content knowledge and thus their pass outcomes. Later when I discuss the findings for Question 3, I describe this evidence. In the meantime, it remains important to consider potential explanations for their significantly poorer initial performances. For example, some

may have not adequately prepared prior to their first attempts but were motivated to amplify their studying efforts when they saw their low initial scores. Others may have performed much more poorly on their initial attempts due to issues such as test anxiety.

2. Research Question 2

   Does examinee performance on different types of items lend support to the pass–fail determinations made about repeat examinees?

   a. How do reused exam items function with respect to distinguishing between different levels of competence among all examinees?
   b. Do repeat examinee subscores, on subtests of items grouped by content area and cognitive complexity level, differ significantly between initial and repeat exam attempt?

Despite some unexpected findings from the preceding section, the findings from Question 2 mostly support observed retest scores and pass–fail outcomes. Based on the results of the item analysis, most reused items were of moderate difficulty and positively discriminated. This suggests that most of the items common across exam attempts had functioned appropriately among all certification examinees, including those who passed their first attempt and did not need to repeat the exam. With respect to item functioning, most items therefore appeared appropriate for assessing repeat examinees. However, item analysis results also identified a number of reused items that poorly functioned among all certification examinees. The results of rescoring repeat attempts with these items removed indicated that the reuse of poorly functioning items may occasionally impact retest pass–fail outcomes, particularly among borderline repeat examinees. The findings from the common item response pattern analyses that I conducted for Question 3(b) moreover suggest these items may have generally hindered repeat examinees'

knowledge remediation efforts. On occasion, the extreme difficulty of some reused items may have actually spurred a studying advantage among repeat examinees. I further explore these findings later in this section of the chapter.

In asking Question 2(b), my goal was to compare how repeat examinees demonstrated knowledge across the full range of content and skills represented on the exam during both attempts. I found that repeat examinees had performed significantly differently across each content area/cognitive complexity subtest regardless of exam attempt, generally performing better on clinical science items as well as on application items. However, neither gain nor loss in subscore was associated with subtest. That is, no subtest significantly lent better to practice effects, knowledge remediation, or any memory advantages or disadvantages. This may be due in part to difficulty with using content area subscore information on initial attempt score reports to correctly identify their relative strengths and weaknesses. That is, subscores are less precise because they are based on small numbers of items and are not separately equated, sometimes resulting in incorrect interpretations by examinees (Julian & Bontempo, 2017; Way & Gialluca, 2017).

The performance differences across subtests on either exam attempt are unsurprising. The Board had written clinical science items and application items to mimic frequently encountered medical scenarios. In that case, everyday clinical practice may have imparted the clinical science knowledge and skills appearing on the exam. This would explain why the repeat examinees in this study performed significantly worse on basic biomedical science items and on recall items. If many of the concepts underlying biomedical science items are indeed less frequently encountered in everyday practice, then poorer performance on these items might signify a

general unpreparedness among the repeat examinees in this study. This would in turn support the observed low retest pass rate.

I had originally considered that cognitive complexity level might relate to improved performance. For example, while attempting to remediate their knowledge, examinees might find it relatively easy to look up the correct answers to recall items, or application items might be easier to remember and research after initial attempts. Wing (1980) found that items that require reasoning are more susceptible to practice effects compared to items that require crystallized knowledge. However, this was not the case in this study. For the repeat examinees in this study, any subscore improvements or losses between exam attempts were generally unrelated to cognitive complexity level, even if performance varied with subtest on each attempt. Both lower and higher cognitive items are susceptible to both positive and negative testing effects (Marsh et al., 2007). Given this, the repeat examinees in the current study may have experienced both memory advantages and memory disadvantages on each of the subtests, hence the insignificant relationship between subtest and improved performance. With each subtest showing similar rates of response persistence and change, this possible explanation is reflected in the findings for Question 3(b).

3. Research Question 3

   Do repeat examinee performances on items to which they have had prior exposure indicate any memory advantages or disadvantages from prior item exposure?

   a. Do repeat examinees score differently on common items compared to unique items?
   b. How do rates of response persistence and response change compare against changes in response time among all repeat examinees?

c. What are the results from 3(b) specifically among passing repeat examinees with score gains beyond measurement error, as identified in Question 1(c)?

Whereas the results for this research question indicate that repeat examinees may have experienced both memory advantages and memory disadvantages due to prior item exposure, prior item exposure ultimately appeared to have a limited effect on retest scores and pass–fail outcomes. Instead, overall content knowledge appeared more associated with overall exam scores and pass–fail outcomes. Repeat examinees had earned, on average, larger score gains on common items than on unique items. However, the average score on common items was significantly lower than on unique items regardless of exam attempt. Furthermore, each individual showed similar ability across common items and unique items on their retest attempts with 95% confidence, lending support to their retest pass–fail outcomes. That is, no individual examinee appeared to pass solely due to prior item exposure advantages or fail solely due to prior item exposure disadvantages. Passing examinees had demonstrated improved knowledge overall, not just on common items. Failing examinees had demonstrated lower ability on unique items, not just on common items.

Regarding unfair score advantages, this study reaffirms the findings from previous studies on medical credentialing repeat examinees. These previous studies have concluded that even though repeat examinee scores often increase on subsequent testing, this generally is not owing to unfair score advantages from prior item exposure (Feinberg et al., 2015; O'Neill et al., 2000; Raymond et al., 2007, 2009; Wood, 2009). After finding that prior item exposure had limited effect on retest outcomes, Raymond et al. (2007) remarked that the time delay of three weeks minimum between exam attempts may have hindered examinees' ability to initially memorize and later recognize common items. In the current study, the time delay was even

greater, ranging from one to three years. Therefore, it is unsurprising that this study's participants did not appear to unfairly benefit from prior item exposure, particularly in light of the lengthy time delays between exam attempts.

The current study revealed better performance on unique exam content than on reused content, which is also consistent with previous studies (Feinberg et al., 2015; Swygert et al., 2010). Through comparison of each examinee's common and unique item subscores, this study has also presented a novel finding regarding the limited effect of memory disadvantages from prior item exposure on retest pass–fail outcomes.

The results for Question 3(b) clarify the memory advantages and disadvantages that repeat examinees may have experienced when re-challenging common multiple choice items. The results altogether indicate that the various advantages and disadvantages stemming from prior item exposure manifested differently between item content areas and different levels of item difficulty, discrimination power, or cognitive complexity. Moreover, persistent knowledge prevailed more often on certain groups of items, whereas other groups of items appeared more susceptible to false knowledge building. Therefore, examinees appeared to differentially demonstrate persistent knowledge, false knowledge, remediated knowledge, and lack of knowledge among repeat examinees on different types of common items.

Previous research has shown that medical licensure and certification examinees tended to reselect response options across both attempts, whether correct or incorrect (Feinberg et al., 2015; O'Neill et al., 2015; Wood, 2009). With correct–correct and incorrect–same incorrect responses accounting for more than half of all common item response patterns in this study, my findings were consistent with these previous studies. Overall, correct–correct patterns were the most frequent, particularly on items that were easy or ideal with respect to difficulty. Though

examinees least frequently demonstrated persistent knowledge on the too difficult items, the accompanying mean decrease in response time when they did so suggests that the high difficulty of certain common items may have made the items more memorable and thereby lent to a studying advantage. Across content area/cognitive complexity subtests, correct–correct was the most frequent response pattern. Therefore, repeat examinees demonstrated persistent knowledge across the range of content areas and skills represented on the exam.

Overall, incorrect–same incorrect was the second most frequent response pattern among common items. Repeat examinees demonstrated persistent false knowledge most frequently on the too difficult and difficult common items. Across the different item discrimination groups, they supplied incorrect–same incorrect patterns at similar rates. However, they appeared to expend more time and thus mental energy when doing so on negatively discriminating items. With respect to the content area/cognitive complexity subtests, both the basic biomedical science recall and application subtests had higher rates of incorrect–same incorrect responses than both clinical science subtests. This is, basic biomedical science content might be more prone to false knowledge building.

Repeat examinees rectified initially incorrect responses about one-fifth of the time. Though they typically did so with a shorter response time, initial and retest response times for incorrect–correct responses were generally longer than those for correct–correct responses. Therefore, prior item exposure appeared more likely to guide knowledge remediation efforts rather than facilitate rote memorization of items and answers. Items of ideal difficulty saw the highest rate of remediation. Despite having the highest rate of incorrect–same incorrect responses, difficult items had the second highest rate of remediated responses. Again, the difficulty of these items may have lent to their memorization and a studying advantage.

Nevertheless, the too difficult items had the lowest rate of remediated responses. Repeat examinees more frequently remediated their responses on positively discriminating items than on negatively discriminating items. Similar rates of incorrect–correct responses occurred between the low and high discrimination item groups. Regarding content area/cognitive complexity subtests, the basic biomedical science recall subtest had a slightly higher rate of incorrect–correct patterns compared to the other subtests. Rates of remediated responses were similar between the clinical science recall and application subtests. My findings contrast with Wing's (1980) assertion that application items lend better to practice effects than recall items.

Several reasons might account for the limited advantages associated with prior item exposure. For example, those retesting are generally of lower ability and perhaps less prepared to accurately memorize and research items or learn from their initial testing experiences (Feinberg et al., 2015; Kulik et al., 1984). The length of a medical credentialing exam and the breadth of the content domain covered may have also impeded memorization (Feinberg et al., 2015; Raymond et al., 2007). When only general performance feedback is provided, repeat examinees may remain largely unaware of what specific content knowledge gaps they need to address prior to retaking the exam (Butler & Roediger, 2008; Feinberg et al., 2015: Raymond et al., 2007). Another potential explanation is that repeat examinees do not think to memorize items to research later because of their confidence in their initial attempts (O'Neill et al., 2015). This includes confidence in initial incorrect responses (O'Neill et al., 2015; Wood; O'Neill et al., 2000; Feinberg et al., 2015). Given that repeat examinees reproduced initial errors more frequently than they corrected them, the results of this study more specifically support the latter possibility.

In addition, many repeat examinee response patterns hinted at memory disadvantages due to prior item exposure. Correct–incorrect patterns occurred most frequently on easy and ideal difficulty items. In these cases, the repetition of false but plausible sounding information about familiar concepts might have facilitated doubt and false recall (Marsh et al., 2007). With lower rates of correct–incorrect response patterns on the too difficult and difficult items, more difficult items may have occasionally related to a studying advantage. On the other hand, the frequency of incorrect–different incorrect responses increased with item difficulty. That is, repeat examinees most frequently demonstrated a persistent lack of knowledge on the more difficult items. Having initially failed, repeat examinees may have been more likely to lack enough knowledge of more advanced, perhaps esoteric, concepts to fully understand and research these items despite prior exposure to them.

On negatively discriminating items, though repeat examinees exhibited more facility in giving initially correct responses, they exhibited more difficulty in rectifying initially incorrect responses. Also, when re-challenging negatively discriminating items, repeat examinees frequently switched from the correct answer to a distractor that had attracted a considerable proportion of all certification examinees. The ambiguity of negatively discriminating items might have fostered false recall.

Comparison of response pattern and time results between the content area/cognitive complexity subtests revealed higher rates of correct–incorrect patterns on the basic biomedical science subtests than on the clinical science subtests. Therefore, basic biomedical science content might be more prone to memory disadvantages such as false recall. Compared to all other subtests, the clinical science application subtest appeared less vulnerable to any memory

disadvantage. Perhaps gaining more clinical work experience in between exam attempts reinforced competence with clinical science application items.

As for the repeat examinees who passed with retest scores beyond 95% probability limits, each examinee demonstrated a higher rate of persistent knowledge on common items compared to the rate for all repeat examinees as a whole. Their respective response patterns and times also indicated that they might have experienced more memory advantages and fewer memory disadvantages due to prior item exposure. In addition, these examinees achieved higher rates of correct responses on unique items compared to all repeat examinees as a whole. Furthermore, the individual *z* tests on common versus unique item subscores indicated that, with 95% confidence, none of these passing examinees performed significantly better on common items than on unique items on their repeat exam attempts. Though examinees who passed with score gains beyond measurement error may have sometimes benefitted from prior item exposure, they appeared to truly remediate their overall content knowledge. They also performed well on unique items across all content areas and cognitive complexity levels reflected on the exam. Therefore, the results from investigating different aspects of their retest performances altogether lend support to their pass outcomes.

B. Practical Implications

The findings from the current study carry several practical implications regarding the reuse of multiple choice items when assessing certification repeat examinees. First, with passing repeat examinees having demonstrated improved overall content knowledge, this study is does not lend support to retest policies that change pass criteria for repeat examinees such as Millman's (1989) ideas to average all exam attempt scores or increase the pass point for repeat examinees. Similarly contradictory to such retest policies, Raymond et al. (2009) concluded that

lower volume credentialing programs might continue to reuse a large proportion of multiple choice items with minimal risk that doing so provides repeat examinees with an unfair score advantage. The findings from the current study support this conclusion, while also qualifying it.

To minimize the potential negative consequences of prior item exposure on repeat examinee outcomes, certification organizations should focus on the quality of the multiple choice items that they reuse on subsequent exam forms. This includes avoiding the reuse of lower quality items that appear more prone to memory advantages or disadvantages among repeat examinees. Based on the findings of this study, medical specialty boards can rest assured that the content area or cognitive complexity level of an item does not significantly relate to any score advantage or disadvantage due to prior item exposure. At the same time, the findings underscore the need to reuse only the items that are both of appropriate difficulty and that positively discriminate among examinees. Exceedingly difficult items and negatively discriminating items might occasionally lend to a studying advantage or thwarted remediation efforts. Common items of ideal difficulty and positive discrimination, on the other hand, appear to better support persistent knowledge and knowledge remediation.

This recommendation should already align with current certification exam development practices within and beyond medicine. Evaluating item quality through item analysis has long been integral to determining which items to reuse on a subsequent exam form (Livingston, 2006). In light of the current research, exam developers must also specifically consider the implications of reusing certain items on an exam form that will be administered to repeat examinees. Certification organizations might be tempted to initially write and later reuse difficult items that, for example, test more emerging knowledge in the field. This study indicates that testing relatively unfamiliar, albeit cutting edge, content can place an undue burden on

repeat examinees in particular. Organizations might also be tempted to reuse an item with less desirable item statistics for content coverage or equating purposes. Based on the results of this study, organizations should instead create a new item around the same content even if that results in losing an anchor item. Assessments intended to measure knowledge can also shape knowledge (e.g., Fazio et al., 2010; Glover, 1989; Marsh et al., 2007; Roediger & Butler, 2011; Roediger & Marsh, 2005). With that, very difficult items as well as flawed items might continually puzzle rather than enlighten repeat examinees. For example, even if medical certification repeat examinees ultimately never pass a specific certification exam, they may still remain licensed to practice and might even become certified in another medical specialty. In general, certification organizations should therefore focus on reusing high quality items that better foster persistent knowledge and knowledge remediation.

The need for ensuring the quality of reused items in order to facilitate classification accuracy applies also to professions outside of medicine. For example, the consequences of credentialing misclassifications in the profession of teaching are similarly borne by both examinees and the public. As I mentioned in the Introduction chapter, false negative classifications might further limit the public's access to practitioners in specialties where there are already physician shortages (e.g., IHS Inc., 2015; Clauser, Margolis, & Case, 2006; Lupu, 2010). With teacher credentialing exams, a higher proportion of repeat examinees are more likely to be racial or ethnic minority applicants than nonminority applicants (National Research Council, 2001). Consequently, false negative classifications made about such repeat examinees might further limit the public's access to diverse educators (Tyler et al., 2011). Reusing high quality items to strengthen the pass–fail determinations made about credentialing repeat

examinees then seems especially crucial for any profession where there is a need to improve public access to qualified practitioners.

C. Significance of the Research

Notwithstanding its limitations, which I discuss later, the current study deepens the existing literature on medical certification repeat examinees and the multiple choice items used to repeatedly assess them. Previous studies have presented results only on repeat examinees as a whole and only on multiple choice items as a whole (Feinberg et al., 2015; O'Neill et al., 2000; Raymond et al., 2007, 2009; O'Neill et al., 2015; Wood, 2009). Moreover, these studies targeted only the possible advantages of prior item exposure, with little consideration of the possible disadvantages. Through this study, I have addressed these gaps in the literature by delving into additional aspects of repeat examinee performance, both at the group level and at the individual level.

Swygert et al. (2010) noted the importance of both group-level and individual-level results when researching medical certification repeat examinees. In response to the previous literature's principal focus on group-level score gains and response patterns, I incorporated individual-level score comparisons and response pattern analyses into this study. With this, my study was able to shed some light on the reuse of multiple choice items with respect to individual score differences and pass–fail determinations.

This study has uniquely explored the relationship between different item characteristics and repeat examinee performance, thereby enhancing interpretation of observed score differences and pass–fail outcomes. For example, through the inclusion of an item analysis, this study has brought to light the bearing that item quality can have on repeat examinee outcomes. Also, Raymond et al. (2007) advocated investigating the relationship between test content and

score gains. The current study addressed the need for such research through an examination of score differences and response patterns on items grouped by content area and cognitive complexity level. Furthermore, this study illustrates the insights that can be achieved through the analysis of common item response patterns and times on items grouped by content area or level of difficulty, discrimination power, or cognitive complexity. Other researchers might adopt this method in their own studies on repeat examinee performance.

As well as aiding interpretation of score differences and pass–fail outcomes, this study's comparison of examinee performance by item type has highlighted the various memory effects that potentially underlie repeat examinee response patterns and times. Rosenfeld et al. (1995) pointed out that the impact of practice effects might vary by item type. However, previous medical credentialing research has paid little attention to this idea, instead concentrating on the relationship between prior item exposure and overall exam score on multiple choice exams (e.g., Feinberg et al., 2015; O'Neill et al., 2015; Raymond et al., 2007, 2009; Wood, 2009). In addition to reaffirming the findings of such previous research, the current research extends those findings by showing how both memory advantages and disadvantages might manifest differently with item type.

In addition to the theoretical contributions of this study, this study makes a practical contribution to retest policy design at medical boards. Medical boards are subject to scrutiny and potential litigation when their constituents start doubting the quality of their certification exams (Knapp & Knapp, 1995; Mehrens, 1995; Millman, 1989). Owing to the use of real data from a single medical specialty board, the key findings of this study might help medical boards develop and refine retest policies that better hold up to scrutiny and litigation. By understanding how repeat examinees perform when re-challenging different types of multiple choice items, boards

can better anticipate how the reuse of certain items might elicit problematic response behaviors from repeat examinees and craft their retest policies accordingly.

D. Limitations of the Research

The current study has generated new insights about the reuse of multiple choice items when assessing medical certification repeat examinees, in turn offering findings that might inform retest policies. Nevertheless, this study has several research limitations. It remains important to acknowledge these limitations and explain how they might have some bearing on the interpretation and practical application of the findings from this study.

First, the sample size of repeat examinees in this study was fairly small ($N = 62$). As I mentioned in the Method chapter, this was a compromise in order to have archival data available on all of the different aspects of repeat examinee performance that I had planned to examine. I had chosen to research this particular sample of repeat examinees over others so that I could more closely investigate repeat examinee performances and response times on different types of items. The resulting sample size, however, presents several limitations. Linacre (1994) identified 50 as a conservative minimum sample size for stable Rasch estimates. I had met this guideline by conducting all Rasch calibrations using the response strings of all certification examinees, not just the repeat examinees, to obtain stable ability estimates. However, in light of the small sample size of repeat examinees in this study, the significance tests I carried out using only repeat examinee scores and subscores had low statistical power.

The fairly small sample size also restricted the factors on retest performance that I was able to analyze. That is, I was unable to divide a sample of this size into the subgroups needed to investigate either of two factors of repeat examinee performance that I mentioned in Chapter II, multiple repeat exam attempts or time delays between initial and repeat exam attempts. Dividing

the sample into subgroups based on number of repeat attempts or on time interval between attempts would have resulted in subgroups too small to produce stable results and therefore meaningful conclusions. Again, the sample size was a compromise for having data available on several aspects of repeat examinee performance. Unfortunately, this means that this study is unable to contribute to the existing research on the relationship between either multiple retakes or time delay and retest performance. For guidance on establishing retest policies, certification organizations might benefit from additional research on these two factors.

Generally speaking, individual certification organizations would be wise to conduct additional research prior to directly adopting the findings of this study. In this study, I investigated repeater behavior for a lower volume medical specialty board. The small sample size, specific testing conditions, and specific scoring procedures featured in this study limit the ability to draw generalizations from any of the findings. For example, to better understand the memory advantages and disadvantages that repeat examinees may have experienced from prior item exposure, I examined repeat examinees' response patterns and times on items grouped by shared characteristics such as item discrimination or content area. Some of the specific characteristics among the common items in this study may have then driven observed response patterns and times. These characteristics may not adequately represent the characteristics of other organizations' reused multiple choice items. Similarly, the sample of repeat examinees in this study may not be representative of repeat examinees in other professions. Consequently, memory advantages and disadvantages from prior item exposure might well manifest differently among other samples of repeat examinees and common items.

Regarding scoring procedures, I had adhered to the scoring procedures that were in operation when this archival dataset had been originally scored. For example, to obtain overall

exam scores, I conducted a separate Rasch analysis for each administration in the five-year period that the dataset covered, equating each administration to the Board's benchmark scale via common-item equating. However, this scoring procedure presents the issue of item parameter drift. Item parameter drift refers to changes in item parameters over time, and it can affect the estimation of examinee ability (Goldstein, 1983). Over the five-year period, the difficulty of some anchor items may have changed due to, for example, advances in medical knowledge. Though a concurrent calibration of all five years of data together would have helped account for item parameter drift, I retained the separate Rasch analyses to maintain consistency with the circumstances under which study participants had repeated the exam and had been scored. Furthermore, I adhered to the original operational displacement threshold of 1.0 logits for unanchoring common items from their anchor values (see Appendix). Some medical boards use a more conservative displacement criterion such as 0.6 logits (e.g., O'Neill, Peabody, Tan, & Du, 2013). Despite this, I adhered to the operational criterion in order to maintain consistency with original scoring conditions. It then remains crucial to point out the potential of item parameter drift to negatively affect the comparability of score performances between the exam years that this study spanned.

Another limitation of this study concerns classification consistency. As I had mentioned in Chapter II, classification consistency indicates the reliability of pass–fail classifications made about examinees across parallel exam administrations and is therefore a priority in the credentialing context (Hambleton & Slater, 1997). However, classification consistency information was absent from the data that I used in this study. This limits the evaluation of exam quality and of the pass–fail determinations made about research participants.

Lastly, given the nature of my chosen research methods, I was only able to base inferences regarding knowledge remediation and memory effects of prior item exposure on the available data, such as response patterns on times. To contextualize repeat examinees' scores when retesting with reused multiple choice items, I chose to analyze archival exam response data using certain quantitative methods. Absent from the archival data were measures of other variables that may have contributed to performances on either exam attempt, such as test anxiety or engagement. This was a non-experimental, descriptive exploration of how repeat examinees might approach common multiple choice items. In the next section, I discuss directions for future explanatory research.

E. Suggestions for Future Research

Despite the theoretical and practical contributions of this study, there remain several areas for future research to deepen the investigation of repeat examinee performances, the consequences of prior item exposure, and the memorial effects of taking multiple choice exams. In this section of the chapter, I describe such research areas. First, I describe future research to resolve unaddressed gaps in the literature. Next, I discuss the directions that future research might take to build upon the theoretical contributions of this study.

As I had mentioned in Chapter II, the available research regarding multiple repeat attempts on medical licensure and certification exams has been scant. Unfortunately, the current study was unable to directly contribute to the research on multiple repeat examinees. This is because the fairly small sample size prevented dividing the study sample into subgroups based on number of repeat attempts. Concerns about allowing multiple exam retakes therefore remain largely unaddressed. On one hand, increased retakes might increase the probability of passing due to measurement error (Clauser et al., 2006; Clauser & Nungester, 2001; Millman, 1989). On

the other hand, increased exposure to multiple choice items might contribute to, maybe even compound, false recall and false knowledge building (Marsh et al., 2007; Roediger & Butler, 2011; Roediger & Marsh, 2005). Future research regarding the performances of multiple repeat examinees, perhaps on different types of items as I had examined in the current study, might address both perspectives.

Though several of the previous studies that I described in Chapter II addressed the factor of time delay on retest performances, the researchers had not investigated how performances might vary across different types of items across different time intervals (Feinberg et al., 2015; O'Neill et al., 2015; Raymond et al., 2007, 2009; Wood, 2009). Again, the fairly small sample size in this study prevented me from analyzing common item response patterns and response times in relation to time delay. Incorporating time interval into an investigation of common item response patterns and time differences might increase understanding of the longer term memory effects that stem from prior exposure to different types of items. In addition, incorporating time delay might provide the insights needed to evaluate retest policies such as imposing a minimum time interval between exam attempts.

Future research on repeat examinees within any high-stakes credentialing context might more closely examine the relationship between initial performances and retest score differences and pass–fail outcomes. Initially, I had planned group-level comparisons of retest performance between examinees who borderline failed versus clearly failed their initial exam attempts. Unfortunately, there happened to be an insufficient number of borderline examinees in the study sample to conduct these comparisons. Additional research that more fully evaluates retest performance with respect to initial performance might help explain how initial levels of knowledge may shape knowledge remediation efforts and memory advantages or disadvantages.

The current study serves as a preliminary exploration of repeat examinee performances and potential memory advantages and disadvantages when re-challenging common items. To extend the theoretical contributions of this exploratory study, researchers might seek to explain repeat examinee score performances and outcomes through additional methods of collecting data about repeat examinee response processes. I had examined archival data consisting of examinee responses, response times, and item content to develop inferences regarding repeat examinees' retest results. Future research might incorporate quantitative or qualitative instruments to gather additional information about additional factors relating to repeat examinee performance. For example, researchers might collect information regarding repeat examinees' respective level of confidence when responding to each common item or their knowledge remediation efforts between exam attempts.

Though the use of such instruments would help generate new insights about repeat testing and the memory effects of prior item exposure, their implementation presents challenges. Quantitative data collection methods such as building inventories to measure additional aspects of repeat examinee performance would facilitate modeling of repeat examinee behaviors and memory effects. Qualitative data collection methods such as interviews and think-aloud protocols with repeat examinees would allow more direct insights into repeat examinee reasoning processes. However, developing such quantitative or qualitative instruments might well require multiple iterations and therefore a lot of resources. In addition, administering them might feel intrusive to or attract scrutiny from repeat examinees within a high-stakes exam context. By honing in on different aspects of repeat examinee performance to contextualize retest outcomes, this study augments the theoretical support necessary to mitigating such research challenges.

F. Conclusion

Medical certification exams are intended to protect the health and safety of the public. The boards that administer these exams must also protect the medical specialty they help govern. They must therefore bear in mind the interests of their constituents, those who practice their medical specialty, when striving to develop appropriate, defensible certification exam policies. With the ultimate goal of producing a study that would be instructive for retest policies, I examined different aspects of repeat examinee performance to contextualize retest outcomes and evaluate the reuse of multiple choice items when assessing medical certification repeat examinees. Though this study yielded some findings that suggest the occurrence of a memory advantage due to prior item exposure among repeat examinees, the other findings indicated that individual score differences and retest pass–fail outcomes were largely related to overall content knowledge. Moreover, indications of the ability to remediate knowledge, the building of false knowledge, and any memory advantages and disadvantages due to prior item exposure varied with category of item difficulty, discrimination power, content area, and cognitive complexity. With respect to the ongoing discussion of how best to assess repeat examinees, this study supports shifting much of the focus from what constitutes an acceptable number of reused items to what constitutes acceptable quality of reused items to administer.

REFERENCES

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

Anderson, L. W., & Krathwohl, D. (Eds.). (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. New York: Longman.

Angoff, W. H. (1984). *Scales, norms, and equivalent scores*. Princeton, NJ: Educational Testing Service. (Reprinted from *Educational measurement* (2nd ed.), pp. 508-600, by R. L. Thorndike, Ed., 1971, Washington, DC: American Council on Education)

Bloom, B., Englehart, M., Furst, E., Hill, W., & Krathwohl, D. (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain*. New York, Toronto: Longmans, Green.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). New York, NY: Routledge.

Boulet, J. R., McKinley, D. W., Whelan, G. P., & Hambleton, R. K. (2003). The effect of task exposure on repeat candidate scores in a high-stakes standardized patient assessment. *Teaching and Learning in Medicine, 15*(4), 227-232. doi:10.1207/S15328015TLM1504_02

Butler, A. C., & Roediger, H. L., III. (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory & Cognition, 36*(3), 604-616. doi:10.3758/MC.36.3.604

Carpenter, S. K. (2009). Cue strength as a moderator of the testing effect: The benefits of elaborative retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*(6), 1563-1569. doi:10.1037/a0017021

Carpenter, S. K., & Delosh, E. L. (2006). Impoverished cue support enhances subsequent retention: Support for the elaborative retrieval explanation of the testing effect. *Memory & Cognition, 34*(2), 268-276. doi:10.3758/BF03193405

Case, S. M., & Swanson, D. B. (2002). *Constructing written test questions for the basic and clinical sciences* (*3rd ed*.) [PDF file]. Philadelphia, PA: National Board of Medical Examiners. Retrieved from http://www.nbme.org/PDF/ItemWriting_2003/2003IWGwhole.pdf

Cianciolo, A. T., Williams, R. G., Klamen, D. L., & Roberts, N. K. (2013). Biomedical knowledge, clinical cognition and diagnostic justification: A structural equation model. *Medical Education, 47*(3), 309-316. doi:10.1111/medu.12096

Clauser, B. E., Margolis, M.J., & Case, S. M. (2006). Testing for licensure and certification in the professions. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 701-731). Westport, CT: Praeger Publishers.

Clauser, B. E., & Nungester, R. J. (2001). Classification accuracy for tests that allow retakes. *Academic Medicine, 76*(10), S108–S110. doi:10.1097/00001888-200110001-00036

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed.). Washington, DC: American Council on Education.

Cronbach, L. J. (1988). Five perspectives on the validity argument. In H. Wainer & H. I. Braun (Eds.), *Test Validity* (pp. 3-17). Hillsdale, NJ: Lawrence Erlbaum Associates.

Downing, S. M. (2002). Construct-irrelevant variance and flawed test questions: Do multiple-choice item writing principles make any difference? *Academic Medicine*, *77*(10), S103-S104. doi:10.1097/00001888-200210001-00032

Downing, S. M. (2006a). Selected-response item formats in test development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 287-301). Mahwah, NJ: Lawrence Erlbaum Associates.

Downing, S. M. (2006b). Twelve steps for effective test development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 3-25). Mahwah, NJ: Lawrence Erlbaum Associates.

Downing, S. M. (2009a). Statistics of testing. In S. M. Downing & R. Yudkowsky (Eds.), *Assessment in health professions education* (pp. 93-117). New York, NY: Routledge.

Downing, S. M. (2009b). Written tests: Constructed-response and selected-response formats. In S. M. Downing & R. Yudkowsky (Eds.), *Assessment in health professions education* (pp. 149-184). New York, NY: Routledge.

Dungan, L. (1996). Examination development. In A. H. Browning, A. C. Bugbee, & M. A. Mullins (Eds.), *Certification: A NOCA handbook* (pp. 1-40). Washington, DC: National Organization for Competency Assurance.

Ebel, R. L, & Frisbie, D. A. (1991). *Essentials of educational measurement* (5th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Fazio, L. K., Agarwal, P. K., Marsh, E. J., & Roediger, H. L. (2010). Memorial consequences of multiple-choice testing on immediate and delayed tests. *Memory & Cognition*, *38*(4), 407–418. doi:10.3758/mc.38.4.407

Feinberg, R. A., Raymond, M. R., & Haist, S. A. (2015). Repeat testing effects on credentialing exams: Are repeaters misinformed or uninformed? *Educational Measurement: Issues and Practice, 34*(1), 34-39. doi:10.1111/emip.12059

Foos, P. W., & Fisher, R. P. (1988). Using tests as learning opportunities. *Journal of Educational Psychology, 80*(2), 179-183. doi:10.1037/0022-0663.80.2.179

Gershon, R. C. (2005). Computer adaptive testing. *Journal of Applied Measurement, 6*(1), 109-127.

Geving, A. M., Webb, S., & Davis, B. (2005). Opportunities for repeat testing: Practice doesn't always make perfect. *Applied H.R.M. Research, 10*(2), 47-56. doi:10.1037/e518612013-432

Glover, J. A. (1989). The "testing" phenomenon: Not gone but nearly forgotten. *Journal of Educational Psychology, 81*(3), 392-399. doi:10.1037//0022-0663.81.3.392

Goldstein, H. (1983). Measuring changes in educational attainment over time: Problems and possibilities. *Journal of Educational Measurement, 20*(4), 369–377. doi: 10.1111/j.1745-3984.1983.tb00214.x

Gunderman, R. B., & Ladowski, J. M. (2013). Inherent limitations of multiple-choice testing. *Academic Radiology, 20*(10), 1319-1321. doi:10.1016/j.acra.2013.04.009

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice, 23*(1), 17-27. doi:10.1111/j.1745-3992.2004.tb00149.x

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15*(3), 309-334.

Haladyna, T. M., & Kramer, G. A. (2004). The validity of subscores for a credentialing test. *Evaluation & the Health Professions, 27*(4), 349-368. doi:10.1177/0163278704270010

Hambleton, R. K., & Slater, S. C. (1997). Reliability of credentialing examinations and the impact of scoring models and standard-setting policies. *Applied Measurement in Education, 10*(1), 19-28. doi:10.1207/s15324818ame1001_2

Hancock, G. R. (1994). Cognitive complexity and the comparability of multiple-choice and constructed-response test formats. *The Journal of Experimental Education, 62*(2), 143-157. doi:10.1080/00220973.1994.9943836

Hasher, L., Goldstein, D., & Toppino, T. (1977). Frequency and the conference of referential validity. *Journal of Verbal Learning and Verbal Behavior, 16,* 107-112. doi: 10.1016/S0022-5371(77)80012-1

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6*(2), 65-70. Retrieved from http://www.jstor.org/stable/4615733

Hutchinson, T. P. (1982). Some theories of performance in multiple choice tests, and their implications for variants of the task. *British journal of mathematical and statistical psychology, 35,* 71-89. doi:10.1111/j.2044-8317.1982.tb00642.x

IHS Inc. (2015). *The complexities of physician supply and demand: Projections from 2013 to 2025.* Retrieved from Association of American Medical Colleges website: https://www.aamc.org/download/426242/data/ihsreportdownload.pdf

Julian, E. R., & Bontempo, B. (2017). Communication with candidates and other stakeholders. In S. Davis-Becker & C. W. Buckendahl (Eds.), *Testing in the professions: Credentialing policies and practice* (pp. 153-177). New York, NY: Routledge.

Kane, M. T. (1982). The validity of licensure examinations. *American Psychologist, 37*(8), 911-918. doi:10.1037//0003-066x.37.8.911

Kane, M. T. (1994a). Validating interpretive arguments for licensure and certification examinations. *Evaluation & the Health Professions, 17*(2), 133-159. doi:10.1177/016327879401700202

Kane, M.T. (1994b). Validating the performance standards associated with passing scores. *Review of Educational Research, 64*(3), 425-461. doi:10.2307/1170678

Kane, M.T. (1996). The precision of measurements. *Applied Measurement in Education, 9*(4), 355-379. doi:10.1207/s15324818ame0904_4

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

Kane, M. T., Kingsbury, C., Colton, D., & Estes, C. (1989). Combining data on criticality and frequency in developing test plans for licensure and certification examinations. *Journal of Educational Measurement, 26*(1), 17-27. doi:10.1111/j.1745-3984.1989.tb00315.x

Knapp, J. E., & Knapp, L. G. (1995). Practice analysis: Building the foundation for validity. In J. C. Impara (Ed.), *Licensure testing: Purposes, procedures and practices* (pp. 93-116). Lincoln, NE: Buros Institute of Mental Measurements.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, linking, and scaling: Methods and practices* (3rd ed). New York, NY: Springer.

Kulik, J. A., Kulik, C. C., & Bangert, R. L. (1984). Effects of practice on aptitude and achievement test scores. *American Educational Research Journal, 21,* 435-447. doi:10.2307/1162453

Lasry, N., Watkins, J., Mazur, E., & Ibrahim, A. (2013). Response times to conceptual questions. *American Journal of Physics*, *81*(9), 703-706. doi: 10.1119/1.4812583

Linacre, J. M. (1994). Sample size and item calibration stability. *Rasch Measurement Transactions, 7*(4), 328.

Linacre, J. M. (2009a). *A user's guide to* Winsteps®*: Rasch-model computer programs*. Beaverton, OR: Winsteps.com.

Linacre, J. M. (2009b). Winsteps® (Version 3.69) [Computer Software]. Beaverton, OR: Winsteps.com.

Little, J. L., & Bjork, E. L. (2015). Optimizing multiple-choice tests as tools for learning. *Memory & Cognition, 43*(1), 14-26. doi:10.3758/s13421-014-0452-8

Little, J. L., Bjork, E. L., Bjork, R. A., & Angello, G. (2012). Multiple-choice tests exonerated, at least of some charges: Fostering test-induced learning and avoiding test-induced forgetting. *Psychological Science, 23*(11), 1337-1344. doi:10.1177/0956797612443370

Livingston, S. A. (2006). Item analysis. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 421-441). Mahwah, NJ: Lawrence Erlbaum Associates.

Lupu, D. (2010). Estimate of current hospice and palliative medicine physician workforce shortage. *Journal of Pain and Symptom Management*, *40*(6), 899-911. doi:10.1016/j.jpainsymman.2010.07.004

Malau-Aduli, B. S., & Zimitat, C. (2012). Peer review improves the quality of MCQ examinations. *Assessment & Evaluation In Higher Education, 37*(8), 919-931. doi:10.1080/02602938.2011.586991

Martinez, M. E. (1999). Cognition and the question of test item format. *Educational Psychologist, 34*(4), 207-218. doi:10.1207/s15326985ep3404_2

Marsh, E. J., Roediger, H. L., III, Bjork, R. A., & Bjork, E. L. (2007). The memorial consequences of multiple-choice testing. *Psychometric Bulletin & Review*, *14*(2), 194-199. doi:10.3758/bf03194051

Matthews-Lopez, J. L., Woo, A., Thiemann, A. J., Jones, P. E., & Gallagher, J. (2015). Test security: A look behinds the scenes. ICE Digest [Online Serial], 3. Retrieved from http://www.credentialingexcellence.org/p/cm/ld/fid=423.

McCallin, R. C. (2016). Test administration. In S. Lane, M. R. Raymond & T. M. Haladyna (Eds.), *Handbook of test development* (2nd ed., pp. 567-584). New York, NY: Routledge.

Mehrens, W. A. (1995). Legal and professional bases for licensure testing. In J. C. Impara (Ed.), *Licensure testing: Purposes, procedures and practices* (pp. 33-58). Lincoln, NE: Buros Institute of Mental Measurements.

Messick, S. (1984). The psychology of educational measurement. *Journal of Educational Measurement, 21,* 215-237. doi:10.1111/j.1745-3984.1984.tb01030.x

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York, NY: Macmillan.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741-749. doi:10.1037/0003-066X.50.9.741

Meyers, J. L., Miller, G. E., & Way, W. D. (2008). Item position and item difficulty change in an IRT-based common item equating design. *Applied Measurement in Education, 22*(1), 38-60. doi:10.1080/08957340802558342

Millman, J. (1989). If at first you don't succeed: Setting passing scores when more than one attempt is permitted. *Educational Researcher, 18*(6), 5-9. doi:10.2307/1176180

Monaghan, W. (2006, July). The facts about subscores [PDF File]. *R&D Connections,* 4, 1-6. Retrieved from https://www.ets.org/Media/Research/pdf/RD_Connections4.pdf

National Research Council (2001). Testing teacher candidates: The role of licensure tests in improving teacher quality. Washington, DC: The National Academies Press. doi:10.17226/10090

O'Neill, T., Lunz, M. E., & Thiede, K. (2000). The impact of receiving the same items on consecutive computer adaptive test administrations. *Journal of Applied Measurement, 1*(2), 131-151.

O'Neill, T., Peabody, M., Tan, R. J. B., & Du, Y. (2013) How much item drift is too much? *Rasch Measurement Transactions, 27*(3), 1423-1424.

O'Neill, T. R., Sun, L., Peabody, M. R., & Royal, K. D. (2015). The impact of repeated exposure to items. *Teaching and Learning in Medicine, 27*(4), 404-409. doi:10.1080/10401334.2015.1077131

Ozuru, Y., Briner, S., Kurby, C. A., & McNamara, D. S. (2013). Comparing comprehension measured by multiple-choice and open-ended questions. *Canadian Journal of Experimental Psychology/Revue Canadienne De Psychologie Expérimentale, 67*(3), 215-227. doi:10.1037/a0032918

Puhan, G., & Liang, L. (2011). Equating subscores using total scaled scores as an anchor. (Research Report RR-11-07). Princeton, NJ: Educational Testing Service.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Chicago, IL: University of Chicago Press.

Raymond, M. R., & Luciw-Dubas, U. A. (2010). The second time around: Accounting for retest effects on oral examinations. *Evaluation & the Health Professions, 33*(3), 386-403. doi:10.1177/0163278710374855

Raymond, M. R., & Neustel, S. (2006). Determining the content of credentialing examinations. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 181-223). Mahwah, NJ: Lawrence Erlbaum Associates.

Raymond, M. R., Neustel, S., & Anderson, D. (2007). Retest effects on identical and parallel forms in certification and licensure testing. *Personnel Psychology, 60*(2), 367-396. doi:10.1111/j.1744-6570.2007.00077.x

Raymond, M. R., Neustel, S., & Anderson, D. (2009). Same-form retest effects on credentialing examinations. *Educational Measurement: Issues and Practice, 28*(2), 19-27. doi:10.1111/j.1745-3992.2009.00144.x

Roediger, H. L., III, & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences, 15*(1), 20-27. doi:10.1016/j.tics.2010.09.003

Roediger, H. L., III, & Marsh, E. J. (2005). The positive and negative consequences of multiple choice testing. *Journal of Experimental Psychology: Learning, Memory and Cognition, 31*(5), 1155-1159. doi:10.1037/0278-7393.31.5.1155

Rosenfeld, M., Tannenbaum, R. J., & Wesley, S. (1995). Policy issues with psychometric implications. In J. C. Impara (Ed.), *Licensure testing: Purposes, procedures, and practices* (pp. 59-87). Lincoln, NE: Buros Institute of Mental Measurements.

Schmeiser, C. B., & Welch, C. J. (2006). Test development. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 324-353). Westport, CT: Praeger Publishers**.**

Smith, I. L., & Hambleton, R. K. (1990). Content validity studies of licensing examinations. *Educational Measurement: Issues and Practice*, *9*(4), 7-10. doi:10.1111/j.1745-3992.1990.tb00385.x

Swygert, K. A., Balog, K. P., & Jobe, A. (2010). The impact of repeat information on examinee performance for a large-scale standardized-patient examination. *Academic Medicine, 85*(9), 1506-1510. doi:10.1097/ACM.0b013e3181eadb25

Tyler, L., Whiting, B., Ferguson, S., Eubanks, S., Steinberg, J., Scatton, L., & Bassett, K. (2011). *Toward increasing teacher diversity: Targeting support and intervention for teacher licensure candidates*. [PDF file]. Washington, DC: National Education Association. Retrieved from https://www.ets.org/s/education_topics/teaching_quality/pdf/support_intervention_teacher_licensure.pdf

Way, W. D., & Gialluca, K. A. (2017). Estimating, interpreting, and maintaining the meaning of test scores. In S. Davis-Becker & C. W. Buckendahl (Eds.), *Testing in the professions: Credentialing policies and practice* (pp. 105-122). New York, NY: Routledge.

Wilson, K. M. (1987). *Patterns of test taking and score change for examinees who repeat the test of English as a foreign language*. Princeton, N.J: Educational Testing Service.

Wing, H. (1980). Practice effects with traditional mental test items. *Applied Psychological Measurement, 4*(2), 141-155. doi:10.1177/014662168000400201

Wood, T. J. (2009). The effect of reused questions on repeat examinees. *Advances in Health Sciences Education: Theory and Practice, 14*(4), 465-473. doi:10.1007/s10459-008-9129-z

Wright, B. D., Linacre, J. M., Gustafson, J. E., & Martin-Löf, P. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*(3), 370.

Wright, B. D., & Stone, M. H. (1979). *Best test design.* Chicago, IL: University of Chicago Social Research.

APPENDIX

Table XIX summarizes the Rasch analyses of the five exam administrations in this study, as well as of the benchmark exam that had preceded these five administrations. The exams were linked to the same benchmark scale through common-item equating.

The upper portion of the table presents fit of each administration's data to the Rasch model. To examine model fit using person and item infit and outfit mean-square statistics. I used the more conservative range of 0.8 to 1.2 to evaluate the infit and outfit mean-square values because of the high-stakes nature of the exam (Wright, Linacre, Gustafson, & Martin-Löf, 1994). Infit or outfit mean-square statistics greater than 1.2 indicate underfit to the model, or excessive variation in person or item response patterns, whereas values less than 0.8 indicate model overfit, or too little variation (Bond & Fox, 2007).

The bottom portion of the table presents item parameter drift information. In keeping with original scoring procedures, I applied a displacement threshold of 1.0 logits. After unanchoring items with substantial drift, all equated exams had at least 20% overlap or at least 30 common items to facilitate common-item equating (Kolen & Brennan, 2014). However, most equated exams had a fairly high proportion of items that could not be used for equating, in turn weakening exam equating (Goldstein, 1983). Advances in medical knowledge over time may have been responsible for item displacement. Also potentially responsible was the change in administration mode from the benchmark exam to the first equated exam, given that even item positioning can affect Rasch item difficulties (Meyers, Miller, & Way, 2008). This would help explain the lower proportions of consistent anchor items on the earliest equated exams. The proportion of consistent anchor items increased over time, perhaps as items from subsequent exam forms became available for equating and earlier anchor items were retired from reuse.

Table XIX

Summary of Benchmark and Equated Exams

Variable	Year 0 Benchmark Exam	Year 1 Equated	Year 2 Equated	Year 3 Equated	Year 4 Equated	Year 5 Equated
Administration mode	Paper-and-pencil	Computer	Computer	Computer	Computer	Computer
			Model fit			
Examinees						
N examinees	310	329	169	174	191	181
n underfitting (%)	27 (8.7%)	24 (7.3%)	10 (5.9%)	11 (6.3%)	12 (6.3%)	9 (5.0%)
n overfitting (%)	25 (8.1%)	19 (5.8%)	5 (3.0%)	3 (1.7%)	2 (1.0%)	0 (0%)
Exam items						
N administered	400	350	300	300	300	300
N scored	364	323	285	287	267	291
n underfitting (%)	12 (3.3%)	36 (11.1%)	10 (3.5%)	23 (8.0%)	26 (9.7%)	15 (5.2%)
n overfitting (%)	26 (7.1%)	39 (12.1%)	25 (8.8%)	23 (8.0%)	32 (12.0%)	11 (3.8%)
			Exam item displacement			
N available anchor items	—	284	88	151	189	160
n displaced (%)	—	169 (59.5%)	46 (52.2%)	50 (33.1%)	47 (24.9%)	29 (18.1%)
n used anchor items (%)	—	115 (40.5%)	42 (47.7%)	101 (66.9%)	142 (75.1%)	131 (81.9%)
% of total scored items	—	35.6%	14.7%	35.2%	53.2%	45.0%

VITA

Lisa A. Reyes
University of Illinois at Chicago
Lreyes24@uic.edu

EDUCATION

2018	University of Illinois at Chicago Ph.D. in Educational Psychology Emphasis: Measurement, Evaluation, Statistics, and Assessment (MESA) Dissertation: "Considerations of Reusing Multiple Choice Items to Assess Medical Certification Repeat Examinees" Advisor: Yue Yin, Ph.D.
2012	University of Illinois at Chicago M.Ed. in Measurement, Evaluation, Statistics, and Assessment (MESA)
2006	University of Chicago B.A. with Honors in Public Policy Studies

RELATED PROFESSIONAL EXPERIENCE

2016 – Present	Psychometrician
2013 – 2016	Research Associate Measurement Incorporated, Chicago, IL
2012 –2013	Certification Services Manager
2012	Certification Services Coordinator Strategic Account Management Association, Chicago, IL
2011 – 2012	Employee Development Coordinator
2008 – 2011	Human Resources Assistant Classified Ventures, LLC (Cars.com, Apartments.com, HomeFinder.com), Chicago, IL
2006 – 2008	Human Resources Coordinator Illinois CPA Society, Chicago, IL

RESEARCH EXPERIENCE

06/2014 – 03/2015 Research Assistant
University of Illinois at Chicago, College of Education

Served as the data analyst for a formative evaluation of science learning assessments, supported by a grant from the U.S. Department of Education through the Chicago Teacher Partnership Program. Performed psychometric and statistical analyses of assessment response data from science education programs at three local universities in order to validate assessment questions and scoring rubrics. Proofread evaluation reports and made revision recommendations.

TEACHING EXPERIENCE

Spring 2018 Graduate Teaching Assistant
University of Illinois at Chicago, Department of Educational Psychology
Educational Measurement

Fall 2015 Graduate Teaching Assistant
University of Illinois at Chicago, Department of Educational Psychology
Essentials of Quantitative Inquiry in Education

PRESENTATIONS

Reyes, L. A. (2018, April). *Score Resolution for Medical Certification Portfolio Exams*. Paper presented at the annual meeting of the National Council on Measurement in Education, New York, NY.

McNaughton, T. & Reyes, L. A. (2016, October). *Developing and Applying Standards in the New Continuous Assessment Framework for Medical Recertification*. Paper presented at the annual Ideas in Testing Research Seminar, Chicago, IL.

Reyes, L. A. (2014, August). *Ways to Examine and Assess Student Performance at an Appropriate Level (Focus on Advanced Programs)*. Workshop presented at the Program Directors Workshop of the American Association of Endodontists, Chicago, IL.

PROFESSIONAL AFFILIATIONS

American Educational Research Association
National Council on Measurement in Education