8-7-2020

Test-wiseness and background knowledge: Their relative contributions to high test performance

Daniel Bennett Roberson

Test-wiseness and background knowledge: Their relative contributions to high test performance

By

Daniel Bennett Roberson

Approved by:

Gary L. Bradshaw (Major Professor)
Andrew F. Jarosz (Committee Member)
Jarrod Moss (Committee Member)
Kevin J. Armstrong (Graduate Coordinator)
Rick Travis (Dean, College of Arts & Sciences)

A Thesis
Submitted to the Faculty of
Mississippi State University
in Partial Fulfillment of the Requirements
for the Degree of Master of Science
in Psychology
in the Department of Psychology

Mississippi State, Mississippi

August 2020

Name: Daniel Bennett Roberson

Date of Degree: August 7, 2020

Institution: Mississippi State University

Major Field: Psychology

Committee Chair: Gary L. Bradshaw

Title of Study: Test-wiseness and background knowledge: their relative contributions to high test performance

Pages in Study 51

Candidate for Degree of Master of Science

When given a multiple-choice test over unfamiliar material, students may score significantly above chance levels. This performance may be explained by prior knowledge of the material or by "test-wiseness," determining the correct answer by using cues present in the test. Participants answered questions from an introductory psychology test-bank in two formats: a question stem with a single alternative and a traditional four alternative multiple-choice, reporting what sources of information they used to answer each question. For the single-alternative condition, participants had an accuracy of 42.2%, 17.2% higher than the base chance of 25%, with an average accuracy of 40.75% for the multiple-choice condition. Participants who stated they had previously learnt the material showed no significant difference in accuracy than those who stated they had guessed. These findings suggest that tests may have inflated scores which reflect test-wiseness and prior knowledge more than formal learning of the test materials.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Multiple-Choice Tests

When it comes to evaluating learning in formal education, the most common testing method used is the multiple-choice test, or MCT. A survey by DiBattista and Kurzawa (2011) found that over half of the undergraduate courses sampled used multiple-choice items in tests, with more frequent use in larger classes (with over 95 students enrolled) and in first- and second-year courses. Research also shows MCTs remain a key assessment method in post-secondary education, even as other assessment methods are used (Mavis, Cole & Hoppe, 2001), with additional research examining the implications of their continued use in professional contexts (Bailey, Mossey, Moroso, Cloutier & Love, 2012).

This widespread adoption of MCTs may lie with advantages of the format. Bacon's (2013) examination of learning outcomes showed that carefully designed MC questions, previously selected and revised based on psychometric properties, took less student time than essays but with equal levels of reliability and validity. Similarly, grading MCTs can take less time through using standardized materials and specialized machines, an appealing advantage for large classes. The objective nature of grading these MC items allows for a perceived lessening in bias, preventing the possibility of a student receiving markedly different grades for their work depending on the reader of the work, such as in Ashburn's (1938) examinations of essay questions. These advantages (among others) encourage support for MC items among students

and instructors, but MC items may have difficulty tapping into higher, application-based knowledge levels (Simkin & Kuechler, 2005).

Another criticism of MCTs derives from the difficulty of constructing items. For example, ambiguous wording of items and their alternatives can further disadvantage students with verbal difficulties (Paxton, 2000). Proper construction of MC items is necessary for proper evaluation of student knowledge. Although texts discussing test theory and item construction are available, surveys of college faculty members indicated that many were unfamiliar with construction procedures, tools, and terms (McDougall, 1997). Examinations of MC items used by faculty (DiBattista & Kurzawa, 2011) found that while one-third showed good discriminatory power, double that amount showed values of .20 or less, failing to meet the benchmark and indicating poor test reliability.

Choosing the Correct Answer

In answering multiple choice questions, students are not simply restricted to a binary outcome of either knowing the answer and thus choosing correctly or not knowing the answer and choosing incorrectly. Rather, students may select an answer based on multiple strategies. Rogers and Bateson (1991) developed a flow-chart model where individuals first attempt to recall knowledge relevant to the question, with success leading to a direct knowledge-derived answer and failure leads to the use of various of test-wiseness strategies to help determine the answer or, at least, help create "educated" guesses by evaluating which alternatives are most likely to be correct.

In brief, Rogers and Bateson (1991) identified 4 different strategies for answering multiple choice questions:

1. Using knowledge about the test content to determine the correct answer.

2. Guessing blindly if no answer is determined.
3. Using test-wiseness strategies to take advantage of the test and its information, resulting in a test-wiseness derived answer.
4. Making an "educated guess" if no answer is determined, making use of a reduced number of possible alternatives.

Guessing

The simplest explanation for choosing a correct answer without knowing it, students may be able to answer questions simply by choosing one of the alternatives at random. In a standard, 4-alternative test, this means a "chance" score would simply be 25%. Under the assumption that students possess some level of knowledge on the material and/or some test-wiseness skills, and that random guessing is unlikely to result in high test performance, students will likely employ other strategies to improve their odds (Downing, 2003); it is more likely that "blind guessing" is only done as a last resort. These blind guesses are distinct from "educated guesses," which themselves are reliant on other strategies to eliminate false alternatives and gauge which alternatives are more likely to be correct.

Learning the Material

Fundamentally, the use of a test, regardless of format, is to gauge a test-taker's knowledge or ability. In educational contexts, tests are given to measure students' knowledge and mastery of specified materials. Material to be learned is presented in the classroom and associated assignments, and therefore equally available for all students. If exams were calibrated to accurately measure student learning, a student who has learnt 50% of the material would score a 50% on an exam.

But prior knowledge and test-wiseness can inflate multiple-choice scores. Assuming the student can correctly identify the correct alternative for 50% of the questions, chance performance on the remaining 50% would inflate their score to 62.5%. Further correct answers may be identified by characteristics inherent to the test itself as well as test-taking skills and extraneous knowledge possessed by students.

Background Knowledge

Another possible attribute that may help in determining an answer is knowledge learned through outside experience. Knowledge learned through outside experience may help in determining an answer is correct. General knowledge allows a question to be answered as it is something that is commonly known or perceived as obvious. While this information may not always allow test-takers to identify the correct answer, as learning the material might, it may allow them to recognize incorrect responses, narrowing down the answer choices available. A student who had not learnt the correct answer in class, but who could use background knowledge to reduce the number of possible alternatives to only two, would improve their 25% score to 50% by such elimination.

Test-wiseness

Test-wiseness represents yet another way where students may select the correct answer without knowing it. Test-wiseness is understood to be the ability of an individual to correctly answer questions independently of possessing knowledge of the correct answer. This is different than the previous two strategies, though overlap in their use is likely common. While background knowledge is also ancillary, test-wiseness focuses on the test and its idiosyncrasies, or on more general test-independent skills like time management. Test-wiseness is demonstrated at the time

of testing, as opposed to learning the material prior to the test. If an individual can discern the correct answer to a question due to the information present in the test itself, this would be test-wiseness.

Test-wiseness can be defined generally as taking advantage of the testing format or situation (Millman, Bishop, & Ebel, 1965) or more specifically as taking advantage of unintentional cues in multiple-choice tests (Gibb, 1964) to improve test performance. This skill is generalizable, meaning that it can be used on tests of varying materials and subjects. However, some tests may be more susceptible to test-wiseness, either due to their format or the material being presented, with the latter interacting with background knowledge. Rogers and Bateson (1991) suggest that the deductive reasoning behind some test-wiseness strategies is dependent on prior knowledge, and test-takers that possess both test-wiseness and prior knowledge will perform better than those who possess only one or the other.

Millman, Bishop, and Ebel (1965) present an extensive taxonomy that details test-wiseness strategies that may be employed by students. Broadly, these can be grouped into two categories: test-independent, where the strategies used are not reliant on the structure of the test or testing situation, and test-dependent, where these factors and testing purpose play a role in strategy use. In the former category, strategies rely on skills such as time-management or error-avoidance, as well as deductive reasoning, which interacts with the previously mentioned background knowledge. As an example, a student may realize they do not know the answer to a difficult question and then decide to set it aside for later and thereby avoid losing valuable time or rashly choosing an answer at random.

In the test-dependent category, the susceptible characteristics of the testing material plays a larger role, as idiosyncrasies and cues are examined and exploited in order to determine the

correct answer, all without knowing the actual material. A student may choose an answer that is notably longer or shorter than other alternatives, or an answer that has more detail. If the participant is familiar with the habits of the test constructor, they may be able to exploit this knowledge, such as by selecting answers based on specific language. If the constructor includes a distractor that opposes the correct answer, by including a negative such as "not," the participant can effectively remove the other answer choices and answer with a 50/50 likelihood of success. Another possible strategy is to consider surrounding questions, which may offer detail that reveals information, such as a question asking for specific information following a general information question. Similarly, recalling specifically emphasized details or information from the constructor may not reveal the correct answer, but offer enough information to determine which items are likely to be correct or incorrect.

Hughes, Salvia, and Bott (1991) found that approximately 75% of tests created by teachers and those provided by publishers contained cued items. Rogers and Bateson's (1991) assessment of school leaving examinations found that 43% to 80% of items were test-wise susceptible, and that students' scores with these susceptible items were significantly higher than those without. The kinds of "educated guesses" brought on by test-wiseness strategies may dramatically increase scores, even when students do not know the correct answer. Instead they can determine what the *incorrect* answers are. While the simple strategy of guessing randomly would result in an average of 25%, these strategies, especially when coupled with helpful background knowledge, can drastically inflate scores for test-takers even though they have not learned the testing material.

All of the Above

While these strategies are distinct, it is likely that multiple strategies are employed (Rogers & Bateson, 1991). If the correct answer is not simply known through prior learning, then students may attempt to use both background knowledge and test-wiseness strategies to help eliminate incorrect alternatives or determine the likelihood of any alternative to be correct or incorrect. If a student can then safely assume an alternative is the only correct choice, then that answer is selected. If multiple alternatives are considered possible, then finally, the student may make an "educated guess" between the lessened number of alternatives. In a standard four-alternative question, if a student is able to eliminate two alternatives as possible choices, then these guesses increase in accuracy from a 25% chance to a 50% chance. When averaged across a number of items in a test, this can result in scores that are significantly higher than a baseline that only accounts for random guessing.

The Validity of Tests

Ultimately, if the purpose of tests is to measure an individual's learning of material, then these extraneous factors decrease the construct validity of tests and hamper their usefulness as metrics of learning in both academic and research contexts. Whether an individual is learning from a lecture, a textbook, or a guided laboratory task, the tests that measure their performance afterwards are subject to internal factors that vary between participants (background knowledge and test-wiseness) and external factors and characteristics of the test itself (e.g. length, number of answer choices, difficulty) which may increase its unreliability.

Individual differences on factors such as both test-wiseness and background knowledge are especially important, as they are inherently unevenly distributed among test-takers. Background knowledge, as a result of education or life experiences, can play a significant role in

separating participants' scores if testing materials contain items that are susceptible. As this background knowledge would not be the focus of the test itself and would be considered ancillary, this uneven distribution can become unfair, and can disadvantage some students while simultaneously benefiting others.

The same can be said for test-wiseness: students who are more skilled at examining the test's structure for cues are able to score more highly than those who are less skilled, even with an equal level of material learnt. Although test-wiseness skills can be taught (Sarnacki, 1979), we must ask ourselves why we are seeking to teach students to perform well even in the absence of learning the material. Presumably we are not training students for a lifetime of multiple-choice tests, but instead educating them with information that will be helpful in their careers and their lives. Coaching test-wiseness skills runs counter to our efforts to educate students.

An important consideration is the role of the test given, which will depend on the context. In a formal education environment, instructors may not give much care to *where* knowledge of the material tested came from, only that the students are familiar enough with it to demonstrate mastery of the concepts taught. In situations such as these, background knowledge may not be considered a detrimental factor for the usefulness of a test. However, separating the factors of learning, prior knowledge, and test-wiseness may still allow for a better understanding of what sources individuals rely on for demonstrating understanding of the test material, which will ultimately provide information beneficial to curricula design and test construction.

Failure to account for these factors may lead to an inaccurate estimate of both what students have learned and what instructors have taught. While high item discriminability may allow for comparative judgments between groups of students (e.g. 'A' students scored 10% higher than 'B' students), these measures may fail to account for the genuine level of knowledge

or mastery either at the individual or group level. If a group demonstrates 10% higher scores than another but has only learnt 30% of the material tested, then this group may be inadvertently given passing scores due to ancillary factors, while still showing a level of separation between other groups. Accounting for factors such as background knowledge not only makes examination fairer between students but can also provide a more accurate *absolute* measurement, which is important in establishing that students have genuinely learnt the necessary materials in a course, and that instructors are effectively teaching said materials.

How Well Can Students Outsmart the Test?

Roberson (2018) examined student performance on publisher's test-bank questions before the students had been exposed to the material in lectures or from their textbook. Participants were presented with test items drawn from two separate test-banks that accompanied introductory psychology textbooks. Two chapters were selected from each test-bank, covering abnormal and social psychology. Even and odd questions from these four tests were then separated into different versions, in order to reduce the number of items which may provide information and answers by covering similar material. Participants were drawn from General Psychology classes at Mississippi State University.

The 222 participants were able to select the correct answer 49.97% of the time, nearly double the chance rate of 25%. In addition, the variability in participant scores was large, ranging from 29% to 79% (Figure 1). Multiple choice items showed an even greater range, between approximately 7% to 98% overall (Figure 2): Some items are evidently highly misleading while others are easy for everyone to guess.

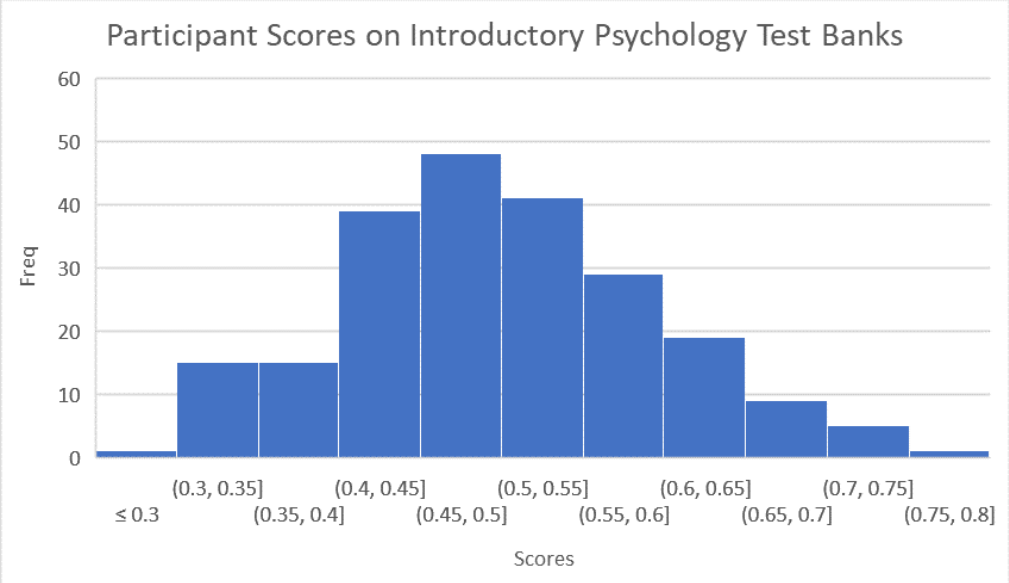

Figure 1. *Previous Overall Participant Scores.*

Frequency of overall participant scores on material from introductory psychology test banks, Roberson (2018)

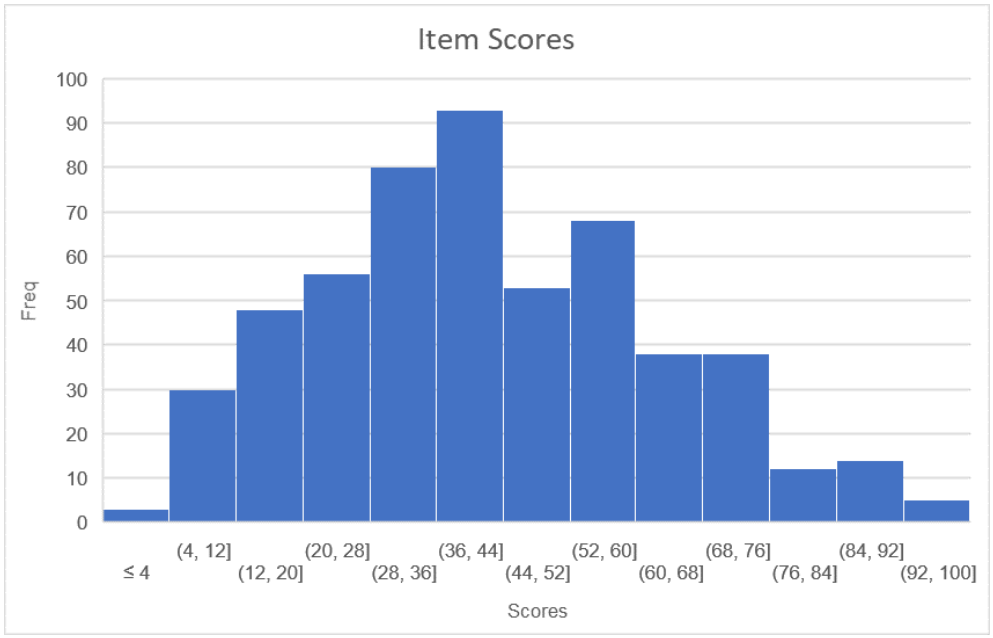

Figure 2. *Previous Overall Item Scores.*

Frequency of overall item scores on material from introductory psychology test banks, Roberson (2018)

In addition to the substantive questions from the test bank, demographic questions regarding previous enrollment in the course and prior exposure to the material were also included. As this experiment continued throughout the full semester, some students reported that they had encountered the tested material in their General Psychology course, either through lectures or textbooks. Roberson (2018) found no significant difference in average scores between individuals who stated that they had encountered the material previously (M = .501, SD = .095) and those who said they had not (M = .501, SD = .098), p = .990. Even more depressingly, participants who indicated they had taken the class previously fared significantly *worse* than other participants, indicating a lack of transfer from their previous class materials to the experimental tests.

From these findings, most students are able to score significantly above a chance level, suggesting the use of test-taking strategies. However, the specific strategies employed are unknown. Students could have employed test-wiseness, background knowledge, or some combination of the two. While these findings suggest that commercially available test-banks may be susceptible to test-taking strategies, the individual strategies' rates of use and effectiveness are still undetermined and require further investigation.

Roberson (2018) also found that items varied drastically in their indices of discriminability. Item discriminability measures the ability for a student's overall performance to be gauged by whether they correctly answer each specific item. Classroom exams are presumably measuring how much of the course content students have learned, and item discriminability measures how well each question can distinguish knowledgeable students from students with poorer mastery of the material. Items with high positive discriminability are indicators of good performance overall, while the items with negative discriminability indicate

that those participants who answered correctly instead had poor overall performance on the test. Participants in Roberson's study had not been exposed to material in class. In this case, participants who did 'well' on the exam presumably had more informal background knowledge, better test-wiseness skills than lower-performing participants, or both, rather than more formal knowledge on the test topics.

This consideration lead to the current study's examination of both highly positive and highly negative discriminability items. Rather than positive, high discriminability items being synonymous with high quality, these items may be just as susceptible to test-wiseness strategies or contain as many salient cues for test-takers as other items. If this is the case, then high discriminability may not be an effective metric by which to judge test items' usefulness in measuring student knowledge when constructing examinations, as those who do well overall may simply be more test-wise than their peers.

Items with negative discriminability are also an important consideration. If these items are answered correctly by students with overall poor performance, they may possess characteristics that make them much more difficult for otherwise well-performing participants. Attributes such as confusingly worded question stems or misleading alternatives may cause participants attempting to make knowledge-based answers fail, as well as disrupt test-wiseness strategy use. By observing performance on these negative discriminability items, it may be possible to see whether participant strategy shifts in response to the item's characteristics, and how effective that might be.

Examinations of test banks' quality often involves experts who attempt to identify vulnerabilities which may allow test-wiseness strategies to succeed (Tarrant et al., 2006; Hansen & Dexter, 1997; Downing, 2002). While these examinations are often capable of noting flaws in

the test, such as misleading items, they often do not account for the background knowledge of the test-takers themselves, which can play a large role in their ability to identify correct and incorrect alternatives. Therefore, the necessary method for examining this ability must be able to independently assess both background knowledge and test-wiseness skills as contributing factors in test performance.

CHAPTER II

EXPERIMENT ONE

This research is designed to answer two research questions. The first: "To what extent do prior knowledge and test-wiseness facilitate selecting correct answers on commercial multiple-choice introductory psychology test banks?" The second: "How do students differ in their ability to employ these strategies to improve their performance on exams?"

To separately assess prior knowledge and test-wiseness, test items can be presented in formats which promote or prohibit the use of one or the other. For prior knowledge, items can be presented with only a single alternative, limiting the availability of cues due to multiple alternatives or viewing other questions. This allows for prior knowledge of the material to be demonstrated through positive confidence judgments on correct alternatives and negative confidence judgments on incorrect alternatives. Conversely, test-wiseness is demonstrated through correctly answering questions only when presented with multiple alternatives at once.

By presenting the questions in both a single-alternative format, which limits available cues, and a traditional multiple-choice format, the extent of test-wiseness may be examined in a paradigm that only requires recognition, not recall. Prior knowledge can be demonstrated even when participants are presented with only a single-alternative for the question, while the availability of cues in a standard, multiple-choice format allows test-wiseness to be employed.

To answer how participants differ from one another in their background knowledge and their ability to use test-wiseness, we can examine the distribution of scores attributable to

differences in their prior knowledge and their test-wiseness. A positive relationship between scores on both tasks would indicate that participants who knew the answer were able to identify the correct choice in both formats, using prior knowledge, while no relationship would point to the multiple-choice cues as necessary for determining a correct answer.

Method

Participants

College undergraduates currently enrolled in an introductory psychology course participated in the experiment in exchange for research credit. A total of 183 students participated, with 8 students' data being removed either due to technical errors or failure to complete the tasks, resulting in 175 participants' data being used in the analysis.

Materials

Thirty questions covering a variety of abnormal and social psychology concepts were used in this experiment. These questions were sourced from test-bank creation software that accompanied two introductory psychology textbooks and were used in previous research (Roberson, 2018.) The 30 items selected were chosen by their discriminability index from the previous research and sorted into two groups: high positive discriminability versus high negative discriminability. An additional question ("What color is an orange?") was included as a manipulation check to discern if participants were on-task. Each of these questions was accompanied by four alternatives: three foils and the correct answer. Note that the set of test questions was developed to minimize cross-talk between items: it was not possible to glean the answer to any given question from material in the remainder of the questions.

This experiment was presented digitally using PsychoPy software (Peirce et al., 2019). All instructions, questions, alternatives, and graphs were presented using a black font on a white background for contrast and legibility.

Procedure

Participants are first presented with instructions and practice problems to familiarize them with the two phases of the experiment. Throughout the experiment, participants are shown question stems alongside either a single alternative or multiple alternatives. In the single-alternative task, only one alternative is shown at a time, with the program cycling through all questions in a random order before the next cycle with the next set of alternatives ("A" alternatives, then "B", and so on.) In the multiple-choice task, questions are presented in a random order, but all possible alternatives are shown simultaneously. Following each question, participants are asked to identify what sources of information they used to either rate the alternative's correctness in the single-alternative task or to select the correct alternative out of the four available in the multiple-choice task (Figure 3).

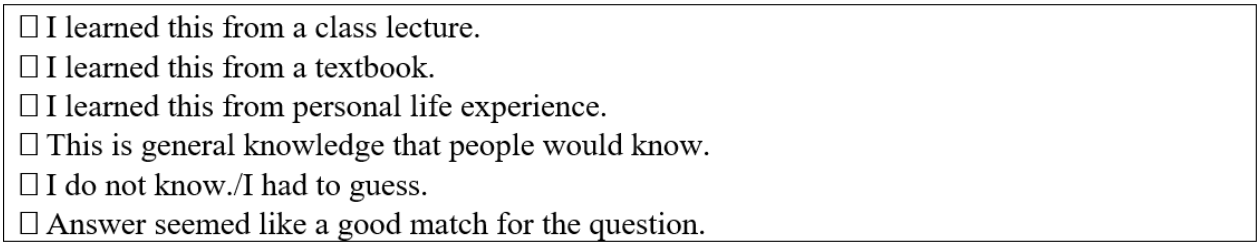

Figure 3. *List of Possible Sources.*

Possible sources participants can select following the questions in both the single-alternative and multiple-choice task. Multiple sources may be selected.

Participants are first presented with a question and a single alternative for that question and asked to rate whether said alternative is correct or incorrect on a five-point scale. They then list any and all sources of information used to arrive at that rating, before moving onto the next question. After all 30 questions have been used, this process repeated using the next set of alternatives, until all four alternatives for all questions had been presented (Figure 4).

A five-point Likert scale was used by participants to rate confidence in whether a displayed alternative was correct or incorrect in the single-alternative condition. This scale ranged from -2 (Absolutely incorrect) to +2 (Absolutely correct) with 0 as a neutral "Unsure." A graphic rating scale was used to measure participants' confidence that their answer was correct following each question in the multiple-choice condition. This scale ranged from 1 (not at all confident) to 100 (completely confident). A list of six possible sources of information (Figure 3) was also presented following these scales. Participants made all selections by using the mouse.

Due to an error in the experiment code, source selections for "general knowledge" erroneously marked "Answer seemed like a good match for the question" as being selected as well in the output files, while marking the latter did not provide any output. This did not affect the display during the experiment. Selection of "seemed" was therefore calculated from vectors where no source was selected, as participants were required to mark at least one source during each attribution. Participants had the option of checking multiple sources; any time they selected "seemed" along with another source was not properly recorded.

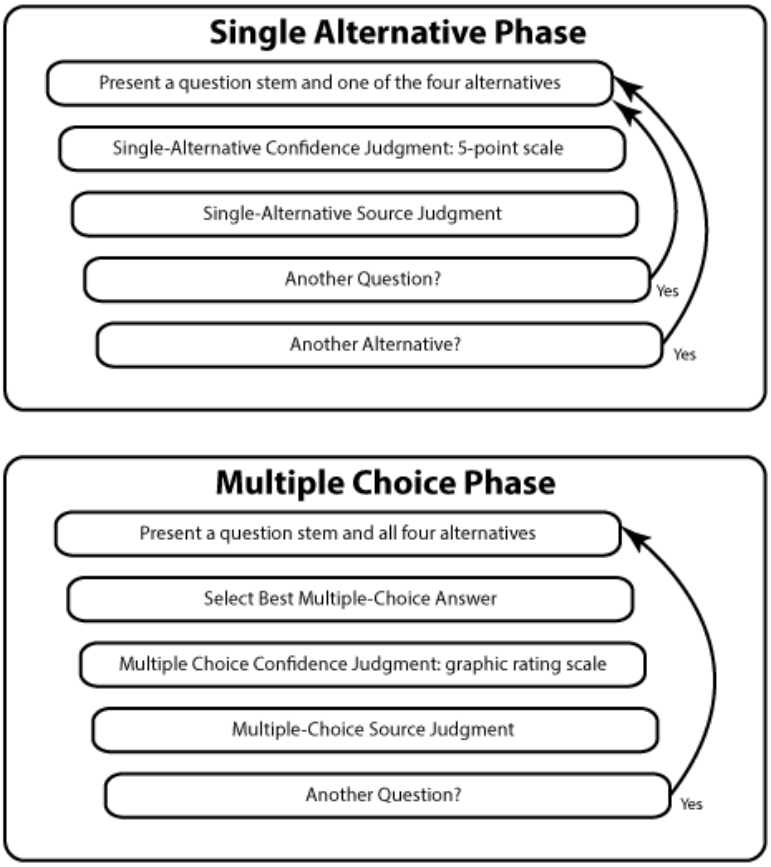

Figure 4. *Sequence of Tasks in Experiment One*

Following the single-alternative phase, participants were then presented with the questions in the traditional four-alternative multiple-choice format and asked to select the correct answer. Next, the participants rated their confidence in their answer using a continuous graphic rating scale (1 – 100), and finally identify the source or sources they used. This process repeated for all 30 questions, and then the experiment ended with the participants being debriefed.

CHAPTER III

RESULTS

To our knowledge, the measurements we are collecting have not been employed previously and little is known about their properties. We need to first evaluate some basic characteristics of these measures. By establishing criteria such as variability between items and participants, further inferential analyses can be made with the knowledge that the resulting statistics are due to the manipulations of the experiment.

Firstly, looking at the variability across items, a large distribution of scores was evident. Figure 5 shows a frequency histogram of the accuracy for the 30 questions on the multiple-choice task, with a range of 16% to 83% and mean score of 40.79%. Of note is the number of scores falling in the 30%-40% range, which is a rate achievable through guessing after removing one false alternative (33%).

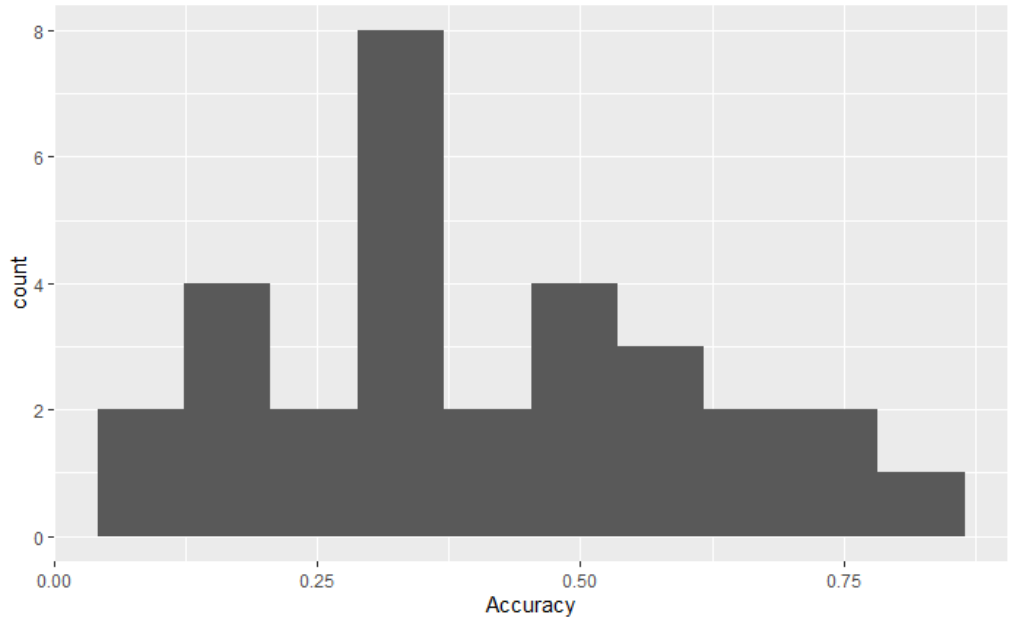

Figure 5. *Frequency Histogram of Multiple-Choice Item Accuracy.*

Secondly, there is the distribution of participants' scores on the multiple-choice task. These scores also show overall accuracy exceeds the 25% chance rate. Participants show a broad range of accuracy similar to our previous research, with some participants scoring well into 60% while others fall below the chance level (Figure 6).

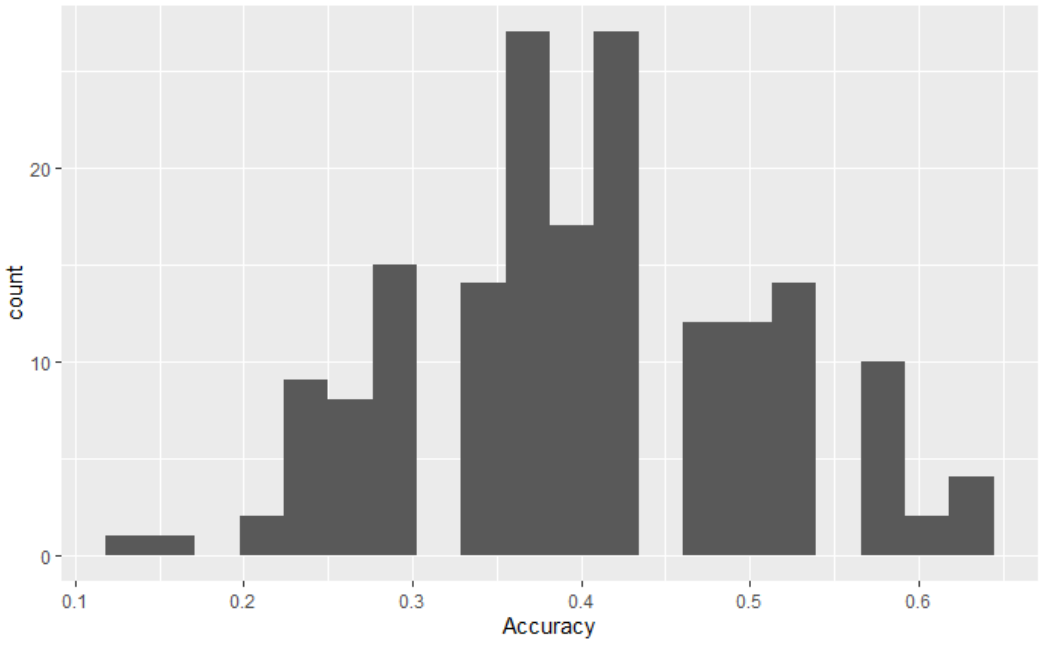

Figure 6. *Frequency Histogram of Multiple-Choice Participant Accuracy*

Independence of Single-Alternative Confidence Judgments

Even though the question stems were presented in randomized order within each block of the single-alternative phase, the alternatives for each question were presented in a fixed order (all A alternatives, then all B alternatives, and so on). Accordingly, there is the possibility that the ratings made on early alternatives could influence the confidence judgments of latter alternatives for that question. For example, recognizing the correct answer was A in the first block logically implies all later selections for the same question are wrong. In such an example where independence is violated, participants who mark A with a +2 rating are predicted to mark the following alternatives with -2 or -1, even though they might have been uncertain about those alternatives if they had been seen earlier. Conversely, participants who marked the first three

alternatives as incorrect would be much more likely to mark the final alternative with +1 or +2, as it is the only option not rejected.

In order to evaluate whether the responses made during the four single-alternative blocks were independent, early ratings were used to predict later ratings. The confidence judgments on C were regressed against the judgments on A and B, while the confidence judgments on D were regressed against the judgments on A, B, and C.

However, these regressions did not reflect the judgments in their presented order. Instead, the maximum and minimum ratings were found for each item and those values were used to predict the rating. This was done in order to use the most confident ratings on each item in terms of both the "most correct" rated alternative and the "least correct" rated alternative. A participant who marked A as +2 and B as -2 would, in a non-independent scenario, but predicted to mark C as -2 as well. The same idea applies for a participant who marked A, B, and C as -2, with D being expected to be +2. The use of these minimum and maximum ratings allows for the correlations between the predictors and predicted rating to be either 1 or -1, depending on the participant's trend of answering across the predictor alternatives.

The very small amount of variance explained by the ratings of previous alternatives on latter alternatives (Alternative C: adjusted $R^2 = .008$; Alternative D: adjusted $R^2 = .034$) strongly suggests that later single-alternative confidence judgments are largely independent of earlier judgments, with only the very large number of observations used (N = 5250) responsible for both regressions' statistical significance ($p < .001$). When predicting C, the minimum ($B = .094$, $\beta = .087$) had a stronger relationship than the maximum ($B = .016$, $\beta = .013$), with the same shown for D, but the minimum ($B = 0.209$, $\beta = .177$) and maximum ($B = -.116$, $\beta = -.084$) predicting opposite directions. This is expected, because if participants were able to answer correctly by

being exposed to the correct alternative early and thus have their following ratings influenced, we would see maximum ratings (the ratings of 2) having a much larger predictive power than the minimum ratings. The range of participants' ratings on alternatives (Figure 7) features a wide spread of answers across the predictive ratings[1].

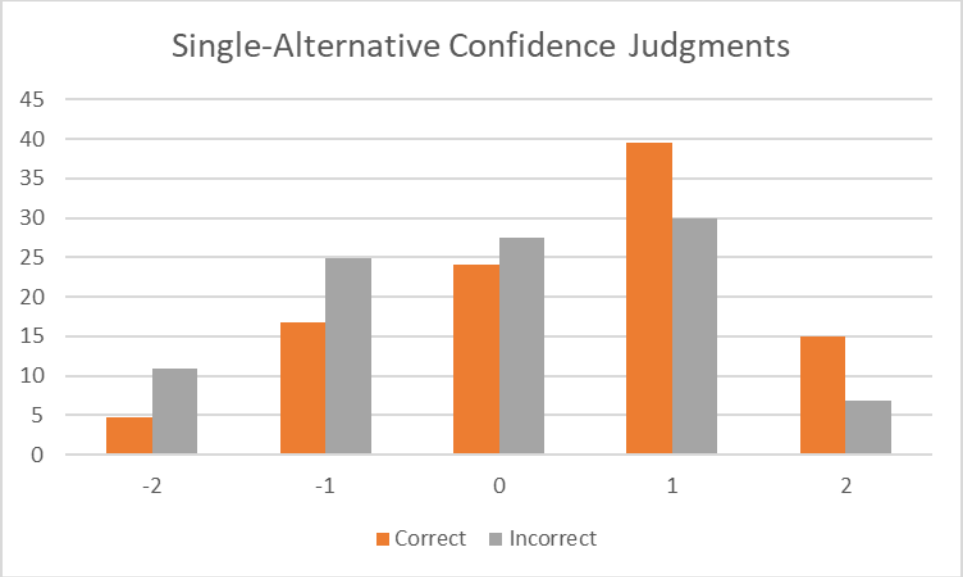

Figure 7. *Single-Alternative Confidence Judgment Ratings*

Participants' overall ratings for correct (orange) and incorrect (gray) alternatives in the single-alternative task.

Additionally, a uniform multinomial distribution of scores on the SACJs was computed to further examine participants' ratings and confirm that they were not affected by the ordering of the alternatives (e.g. giving a +2 rating to the first alternative and then -2 to all subsequent

[1] As a rating of zero is "Unsure", it does not help predict future ratings by indicating the participant knows something about the correct/incorrect answers. Therefore, cases where the participant is unsure and answers with zeros may be removed from the regression, with the remaining cases being only those with possible predictions from participants. However, these new regressions for C ($N = 3235$, adjusted $R^2 = .011$) and D ($N = 2709$, adjusted $R^2 = .041$) show little improvement over the standard regressions.

alternatives). These ratings were created by summing up the total confidence ratings on each alternative of an item, reversing the scores for the incorrect alternatives (-2 would become a +2, and so on.) Comparing them to the experimental results (Figure 8), only minor deviations are shown. This demonstrates that, rather than a significantly larger proportion of higher scores due to early identification of the correct alternative, the data are close to the model of complete independence of ratings.

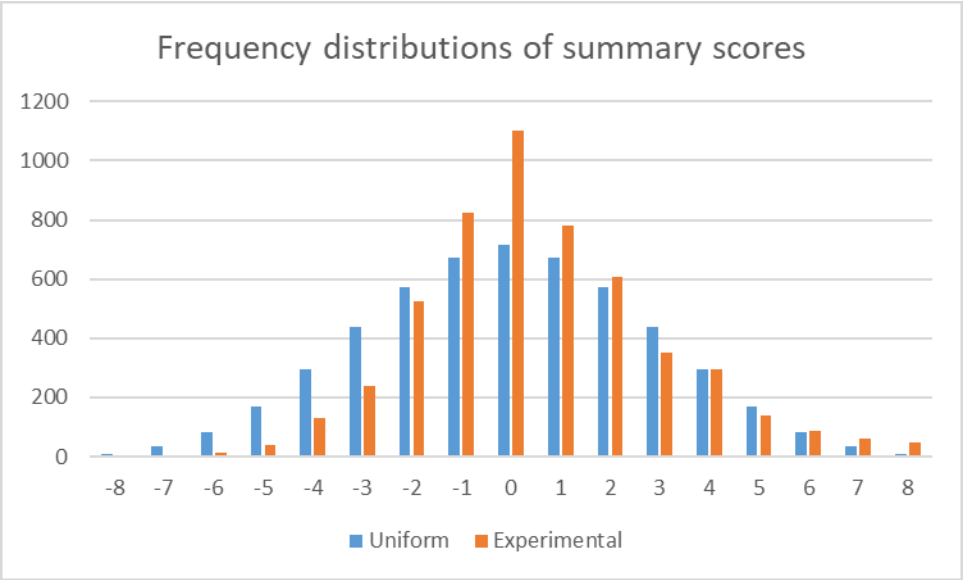

Figure 8. *Comparison of Frequency of SACJ Scores*

This figure compares participants' single-alternative confidence judgments (orange) to a uniform multinomial distribution (blue).

Single-Alternative Confidence Judgments – Effects of Item and Participant Characteristics

Immediately following each single-alternative confidence judgment, participants were asked to make a corresponding single-alternative source attribution (SASA). Participants were asked to select any and all sources of information they used when answering the SACJs (Figure 3). Through analyzing the rates of selection for these sources, we may identify uses of previous

knowledge or test-wiseness strategies and how participants' sources change as a result of differences in item characteristics, such as discriminability.

With six choices for attribution, a vector can be created, and the participant's choices across the four alternatives can be summed together. As an example, a participant who indicated that they had learnt the material in class for all four alternatives would have a vector with 4 for that source and 0 for the other sources. Four is the maximum value possible for a source in the vector and zero is the minimum. A participant who was inconsistent in their selection may select a different source for each alternative, resulting in a vector where four sources have a value of 1, as no source is selected more than once.

For the SASAs, a series of four analyses were performed on these vectors, which compared the frequency of source attributions as they may be affected by other variables including items, participant accuracy, discriminability index, and item correctness. If differences are shown in source attributions, this suggests that the source of data is useful in understanding performance on the task. In a degenerate case, everyone might universally report 'Don't know/had to guess' on all items, in which case the data will not be useful in later analyses.

The first independent variable examined are the items and how they affect the sources participants select. As predicted, there was a significance difference between the sources used based on the item presented, $\chi^2(145) = 9042.9$, $p < .001$. This suggests the data is useful in efforts to relate source judgments and performance.

The second analysis examined differences of sources between correct and incorrect alternatives. Looking at the table of observed source attributions (Table 1), "guess" is cited as the source much more frequently than others, for both correct and incorrect alternatives. The percentage selecting 'guess' for incorrect answers, however, is greater than for correct.

Participants may struggle to cite a specific source of information when stating an alternative is false but be more specific when they are affirming that an alternative is correct.

Table 1

Single-Alternative Source Judgments by (In)Correct Alternatives in Percentage

	Class	Textbook	Personal	General	Guess	Seemed
Incorrect Alternatives	11.82	5.00	6.92	14.93	42.69	18.63
Correct Alternatives	14.00	6.99	8.97	16.40	35.60	18.04

The third analysis compared the frequencies of sources used between the two item discriminability groups (positive and negative item discriminability). Based on the table of observed results (Table 2), the largest difference appears to be an increase in guessing for negative discriminability items, while positive discriminability items have more attributions to general knowledge or test-wiseness ("The answer seems to be a good match for the question.") This result matches the prediction that high discriminability items either require prior knowledge of the material to select the correct answer (hence the increased number of citations for prior knowledge, like "class") or higher test-wiseness in order to determine the answer (higher "seemed" rate). Meanwhile, negative discriminability items are poor predictors of overall performance, likely by obfuscating the correct answer with difficult to parse questions or particularly misleading distractor alternatives.

Table 2

Single-Alternative Source Judgments by Discriminability Group in Percentages

	Class	Textbook	Personal	General	Guess	Seemed
Positive Discriminability	6.78	2.92	3.51	8.51	18.15	10.16
Negative Discriminability	5.60	2.60	3.94	6.80	22.71	8.33

The fourth analysis compared source attributions between high-scoring and low-scoring participants, using their scores on the single-alternative task. The participants were split into four ability groups by quartiles, with Quartile 4 being the highest scoring group. Looking at the table of observed results (Table 3), it appears that the sources used by participants had large differences in their rates of use, with guessing accounting for approximately 40% of sources listed. In comparison, participants cited using information learnt from a class or textbook much less frequently than guessing or stating that the answer chosen "seemed" correct, indicating a test-wise judgment. Surprisingly, the rates of selection for these sources did not differ greatly across the quartiles. With "guess" and "seemed" being the most often selected sources, this suggests that accurately rating an alternative as correct or incorrect is dependent on the participant's ability to make well-reasoned "educated guesses," rather than their ability to recall previously learnt information from other sources.

Table 3

Single-Alternative Source Judgments by Participant Scores (Quartiles) in Percentages

	Class	Textbook	Personal	General	Guess	Seemed
Quartile 1	13.59	5.31	6.49	14.12	41.70	18.78
Quartile 2	9.33	5.93	6.50	13.24	43.20	21.80
Quartile 3	11.58	4.74	8.64	16.72	42.04	16.28
Quartile 4	14.99	6.08	8.12	17.07	36.58	17.15

In total, these analyses highlight differences across the aggregate source attributions, helping to further define the item and participant characteristics that lead to selection of particular sources.

Multiple-Choice Accuracy

Accuracy on multiple-choice questions is binary, either correct or incorrect, in contrast to the SACJs scale of possible scores. Participants were presented with a question, along with its four possible alternatives, and asked to select the correct answer (Figure 9). Correctness on multiple choice questions was analyzed using a gamma correlation with the discriminability index predicting accuracy, grouped by participant. These participant gammas were then used in a one-sample t-test to compare them with a baseline of 0, to determine if there was a significant effect of item discriminability on participant accuracy.

In addition to a belief that strong _____ factors may influence schizophrenia, researchers have found evidence that exposure to viruses _____ may also increase the risk of schizophrenia.
☐ **A.** cultural; in the womb ☐ **B.** psychic; during the second year of life ☐ **C.** biological; prenatally ☐ **D.** environmental; during late adolescence

Figure 9. *Example of Multiple-Choice Question and Alternatives*

There was a significant difference found between the baseline of 0 and the participants' gamma correlations ($M = .261$) between their accuracy and the items' discriminability, $t(174) = 14.352$, $p < .001$. Based on the average positive value of the gamma correlations, the overall accuracy for items is higher for the positive discriminability items than the negative discriminability items. This provides support for the idea that high-discriminability is, of course, a good metric for predicting overall performance on the test, but that it is not (considering the unfamiliarity of the material to participants and their sources used) a metric which guarantees those who succeed do so through having learnt the test material.

Multiple-Choice Source Attributions – Effects of Item and Participant Characteristics

There are also the multiple-choice source attributions (MCSAs) given after the participant selects an alternative and provides their confidence rating, identical to those used in the SASAs (Figure 3). These are not aggregated across item alternatives as they were for the single-alternative task, as this phase of the experiment has each item only appearing once.

The first analysis compared the frequencies of source attributions based on the items. The source attributions given were significantly different across the items presented to participants, $\chi^2(145) = 2656.8$, $p < .001$.

The second analysis examined the two item discriminability groups (positive and negative discriminability). Looking at the table of observed results (Table 4), it appears that positive-discriminability items have higher rates of classes or textbooks being cited as a source, but also general knowledge and test-wiseness. The greater use of both prior knowledge and test-wiseness is expected, while the increase in proportion for test-wiseness for this task over the single-alternative task (Table 2), points to the increased availability of cues due to the multiple-choice format resulting in greater test-wiseness use, if not efficacy.

Table 4

Multiple-Choice Source Judgments by Discriminability Groups in Percentages

	Class	Textbook	Personal	General	Guess	Seemed
Positive Discriminability	7.67	3.34	4.08	8.60	17.09	9.36
Negative Discriminability	5.72	2.71	4.84	6.45	22.83	7.30

The third analysis examined participants divided into quartile groups based on participants' overall accuracy on the multiple-choice questions. With the table of observations (Table 5), it seems that the lowest performers (Quartile 1) are less likely to indicate learning the information from class or personal experience than the highest performers (Quartile 4).

Table 5

Multiple-Choice Source Judgments by Participant Accuracy (Quartiles)

	Class	Textbook	Personal	General	Guess	Seemed
Quartile 1	10.72	6.66	7.03	15.61	40.35	19.63
Quartile 2	15.31	4.82	7.64	13.14	43.69	15.39
Quartile 3	13.97	6.49	10.67	15.39	38.60	14.89
Quartile 4	13.46	6.19	10.20	16.08	37.12	16.95

These analyses not only highlight the levels of consistency in alternative selection and how they are affected by these participant and item characteristics, but also provide additional information on how test-wiseness can improve participants' ability to correctly answer questions, even with minor changes such as presentation of multiple alternatives.

Given that the internal consistency of our measures has been established, we can begin answering the primary research questions regarding test quality and student performance. These questions examine: 1) the extent to which students exploit test-wiseness and prior knowledge, 2) whether high-performing students have greater meta-awareness of their performance, and 3) how item discriminability affects confidence of students of different performance levels.

CHAPTER IV

RESEARCH QUESTIONS

Question 1 – "To What Extent Do Students Exploit Test-Wiseness and Prior Knowledge to Answer Questions?"

In this experiment, participants were subjectively judged to be using test-wiseness to answer multiple-choice questions based on their source attributions. While participants can state that they used multiple sources to answer a question, test-wiseness use was defined by two criteria: they *do not* select a source indicating previously learning the material (e.g. "I learnt this from a textbook.") and they *do* select the alternative chosen "seems to match the question." These source attributions help to sort participants' answers into groups: 1) Retrieving prior knowledge from outside of academia (citing personal experience or general knowledge); 2) Learning through formal education (citing classes or textbooks); 3) Employing test-wiseness (defined above); and 4) Random guessing (citing having to guess).

The 6 original source attributions were collapsed into the four categories above. One important consideration for the following analyses is the exclusivity of these categories. Participants may, for any question, select multiple sources as being used to provide an answer. Having multiple categories being listed for a single question, however, dilutes their independence and may cause further issues regarding the assumptions necessary for the analyses. Therefore, it is important to examine just how often multiple sources were selected, and how these multiple selections are distributed.

Of the 5225 total responses for the multiple-choice task, 89% had participants indicate only one source was used, with 9% having two, and the remaining 2% having three to five. For the questions where two sources were cited, most were within categories (e.g. class and textbook or general knowledge and personal experience) and could be combined without issue. In total, 95.87% of the total source attributions for the multiple-choice task were category exclusive. As the remaining 4.13% are spread across multiple conflicting categories, they have been discarded for the remaining analyses.

When looking at the table of observed attributions for correct and incorrect answers (Table 6), the largest difference observed is the high frequency of guessing for incorrect answers. Participants who reported they were guessing had a high probability of having selected the wrong answer. Curiously, students who did report exposure to the material in a formal learning setting, and thus presumably 'knew' the correct answer, had no better chance of selecting the correct answer than students who employed background knowledge or test-wiseness cue.

Table 6

Multiple-Choice Source Judgment Groups by (In)Correct Answers

	Formal Knowledge	Informal Knowledge	Guess	Test-wiseness
Incorrect Answers	378	514	1497	577
Correct Answers	360	511	709	407

It is also possible to approximate the proportion of prior knowledge used in the SACJs through observing the difference between chance performance (25%) and their overall

performance. This overall performance is calculated using the participants' ratings for the correct and incorrect alternatives of an item (Figure 7). As these ratings are not just a binary choice, a "knowledge-based recall" for an item can be operationalized in multiple ways. As this task's role is to highlight the role of prior knowledge by eliminating test-wiseness, the operationalization should be focused on showing participant performance due to only recalling previously learnt information, rather than using test-wiseness cues or strategies, such as comparing the alternatives against one another.

The simplest way to rate an item as correctly answered would be to compute a synthetic recognition score by comparing the rating for the correct alternative against all of the incorrect alternatives. If the correct alternative receives a higher rating than all incorrect items, we judge they successfully selected the correct answer. Those who provide equal maximum ratings for correct and one or more incorrect alternatives would be considered to have answered correctly half the time, while those who rate the incorrect alternative as more likely to be correct than the actual correct answered would be considered to have missed the question. Using this method for all participants and their items, we obtained an average score of 42.2% for participants.

However, as this method of scoring does not factor in the correct rejection of false alternatives, it may inflate participants' scores in the single-alternative task. Participants may rate both the correct and incorrect alternatives as "Likely Correct," leading to an overall answer that is more likely an educated guess based on test-wiseness than a definitive recall of prior knowledge. While choosing the alternative that is "most correct" is a valid option during test-taking for narrowing down answer choices, correctly recalling previously learnt information is expected to not only inform participants of what the right answer is, but to more definitively provide information to discard false alternatives.

Participants who used prior knowledge to correctly answer the questions in the single-alternative task were expected to answer the same questions correctly in the multiple-choice task. This is due to their use of recalled information, rather than the use of internal cues which facilitate test-wiseness strategies. However, using the aforementioned method of comparing the highest rating for correct and incorrect alternatives, participants had an average score on the traditional multiple-choice format of only 36.24% on those items judged to be 'correct' using the synthetic score based on SACJ's. Therefore, the average SACJ score of 42.2% appears to overstate how many items participants were able to recall on their combined confidence judgments.

A more selective method for rating participants' responses is to define a correct response for an item that has a positive confidence for the correct alternative (1 or 2) and a less-than-one judgment (-2, -1, 0) for all incorrect alternatives. This method treats the rating of "Likely Correct" or "Absolutely Correct" for the correct alternative as equivalent to their selection in a multiple-choice context, with neutral or negative ratings for the incorrect alternatives as correct rejections. In summary, this would demonstrate prior knowledge use to answer these questions, as the participants are able to clearly separate the correct alternative from the incorrect ones.

With this operationalization, participants had a 13.1% recall rate overall, a much lower rate than suggested by the lenient scoring method. However, this did demonstrate a much higher rate of accuracy for the multiple-choice items which they had answered correctly in the single-alternative phase. Participants' average score on these items was 69.81%, suggesting these stricter results are likely demonstrating the contribution of prior knowledge on both tasks.

If we further insist on a rating of 2, the maximum rating of "Absolutely Correct", for the correct alternative and retain the same range (-2, -1, 0) for incorrect alternatives, the synthetic

recall rate for the single-alternative task is only 4.3%. Finally, if we define a correct recall as a (1, 2) for correct and (-2, -1) for all incorrect alternatives, meaning that no alternative is rated with "Unsure," recall is 6.7%.

Overall, participants demonstrated the ability to identify what alternative is "most likely" (rating the correct alternative above the incorrect alternatives) at a rate 17.2% above a chance 25% rate of randomly selecting one alternative as the most correct. This rate is similar to what has been shown previously in regards to participants' ability to select answers through "educated guesses" on multiple-choice questions (Roberson, 2018). This suggests that participants may be able to employ test-wiseness strategies, even with only a question stem and alternative, in order to gauge alternatives' relative likelihood of being correct. However, participants were much less able to both identify the correct alternative for a single-item and reject the incorrect alternatives, leading to prior-knowledge-based recall rates of no greater than 13.1%.

Next, we can compare this derived accuracy measure against the final performance on the complete multiple-choice items. The average of participants' scores on the multiple-choice task was 40.75%. Subtracting a baseline 25% achievable through randomly guessing, the average multiple-choice score is 15.75%, a higher rate of performance than the 13.1% rate estimated by the derived single-alternative method. This increase over baseline guessing was expected, as we anticipated that availability of test-wiseness cues in the full multiple-choice items would enable participants to more frequently identify the correct alternative and eliminate other alternatives.

However, it is important to consider the role of prior knowledge responses on the multiple-choice task, as well. Participants still showed an approximate 70% accuracy for multiple-choice items that they answered correctly using the stringent scoring method on the single-alternative task. Therefore, the most conservative estimate of the role of test-wiseness on

the MC would consider only those items that participants failed to answer correctly in the single-alternative task, based on the idea that these items are not being answered by prior knowledge and must be correctly answered using only the available test-wiseness cues. When looking at participant accuracy for only these items, with subtraction of 25% to account for random guessing, overall accuracy on the multiple-choice task was 11.40%. In other words, participants who did not demonstrate prior knowledge of the test material that allowed them to identify the correct answer to a question were not just randomly guessing for answers, but used the available information in the items to increase their score by 11.40% overall.

Question 2 – "How Do Students Differ in Their Ability to Employ These Strategies to Improve Their Performance on Exams?"

Six possible sources can be identified for each item, but these can be collapsed into four distinct factors: formal learning, general knowledge, test-wiseness, and guessing. We can use these factors to see which (if any) predict overall test performance. A multiple regression was performed that predicts overall score in the multiple-choice task for each individual as a function of their source selections (number of formal selections, number of general knowledge selections, number of test-wiseness selections, and the number of guess selections). An obvious hypothesis: performance will be facilitated by general knowledge and test-wiseness, while guessing should produce lower levels of performance. Formal learning should also be associated with high levels of performance, though this was a less frequent choice in the data.

This regression failed to demonstrate any of these factors were significant predictors, $F(4, 170) = .156$, $p = .960$, with each source providing little explanation of the variance (Table 7). Thus, students who reported frequently guessing did not perform any worse than students who reported using formal knowledge to answer a larger share of questions.

Table 7

Coefficients of Accuracy and Sources Multiple Regression

	B	*β*	*t*	*p*
Formal Knowledge	.013	.018	.117	.907
Informal Knowledge	.040	.057	.389	.698
Guess	-.005	-.010	-.052	.958
Test-wiseness	.007	.012	.074	.941

This result is surprising and disconcerting. Participants' success on these multiple-choice test questions are certainly higher than what would be expected through random guessing, but there is no obvious separation for those who reported knowing more and those who reported guessing or using test-wiseness. This suggests that participants have difficulty accurately reporting the source of their knowledge. A possibility is that this analysis, which looks across all items, may be failing to discern differences which arise from item discriminability, which was previously shown to significantly affect which sources are used.

Question 2a – "For Positive- and Negative-Discriminability Items, How Confident Are Students of Different Levels About Their Performance?"

Building off of the previous question, a second analysis examined the differences in test-wiseness and prior knowledge demands of questions based on discriminability. Items with positive discriminability may be easily answered with the prerequisite prior knowledge or present their question and alternatives in a way that provides information to test-wise participants. In contrast, items with negative discriminability indicate poor overall test performance when answering them correctly. Such items may use misleading distractors as alternatives or promote misuse of effective test-wiseness strategies. They may also have lessened predictive power by presenting easily dismissed alternatives which any test-taker can identify as

incorrect, regardless of their knowledge over the material or overall test-wiseness. We predict that the characteristics and strategy demands of the items are related to their overall discriminability, and that this discriminability can significantly affect the confidence and ability of participants to answer correctly.

A linear mixed-effects model was used to predict participant's multiple-choice confidence ratings as a result of the interaction between their accuracy and the item's discriminability as a fixed effect, with accuracy also being implemented as random effects grouped by participant and item, in order to account for inter-factor variability caused by repeated measurements for the items and participants. The summary of the fixed effects is shown in table 8.

Table 8

LMEM Fixed Effects for Multiple-Choice Confidence

	Estimate (confidence)	*SE*	*df*	*t*	*p*
(Intercept)	54.77	2.64	36.17	20.72	< .001
Accuracy	4.86	1.63	28.74	2.99	.006
Discriminability	6.71	5.37	28.25	1.25	.222
Accuracy: Discriminability	6.60	3.49	27.75	1.89	.069

Multiple-choice confidence predicted with accuracy, discriminability, and their interaction as fixed effects, and accuracy group by participant and item as random effects. Data were analyzed using R through the lmerTest package (Kuznetsova et al., 2016)

Accuracy is shown to be a significant predictor of confidence in the model. However, comparing this model to a null model, a marginal R^2 of .046 shows little variance in confidence explained by participant accuracy and item discriminability. While item discriminability has been shown previously to affect the sources participants say they used to answer, participants

may not be sure enough in their answers to always report high confidence on their accurate answers or on high discriminability items they may know the answer to.

CHAPTER V

DISCUSSION

A new paradigm was developed to explore the sources of knowledge used by undergraduates to select the correct answer for multiple-choice questions in introductory psychology. This method contrasts recognition judgments based on the question stem and a single alternative with traditional multiple-choice recognition judgments. Supplemental data was obtained by asking participants to indicate, item-by-item, the source of their knowledge.

With a simple method of operationalizing the single-alternative judgments based on the rating for the correct alternative, we found that participants had a rate of correct response of 42.2%, 17% above chance-level. However, this method fails to definitively separate ratings for alternatives into a recognition of the correct alternative and rejection of the incorrect alternatives. Using a stricter operationalization, with correct answers being positive ratings (1 or 2) for the correct alternative and negative/unsure ratings (0, -1, -2) for all incorrect alternatives, the participants had a correct response rate of 13.1%. In contrast, performance on the traditional multiple-choice format, subtracting a base-rate of 25%, was higher, at 15.75%, with a conservative estimate which accounts for prior knowledge-based answers showing a similar increase of 11.40%.

Rather than more definitive identification of what answer choices are right and which are wrong, student success may rely heavily on systematic removal of unlikely alternatives through the use of test-wiseness cues, including the alternatives themselves. Typical classroom tests may

additionally inflate student performance by choosing questions that tap into general knowledge or that offer clues in one question that may help to answer another. Considerable variability was observed both in item difficulty and in prior knowledge. These findings suggest that student performance in a classroom may reflect general prior knowledge and test-wiseness more heavily than classroom learning and recall of that information on tests.

Source judgments showed little predictive ability in accuracy, other than when participants reported they guessed at the answer and tended to pick the wrong alternative at a higher frequency than when employing formal knowledge, general background knowledge, or test-wiseness. The lack of any advantage for having reported that participants learned the material in a classroom or in a textbook implies students are not learning the material well-enough to select the correct answer more than 42% of the time, a disturbing outcome consistent with prior research (Roberson, 2018).

The items presented to the participants varied in terms of their discriminability. Positive-discriminability items led to more attributions to prior knowledge and test-wiseness, with negative-discriminability items having more guesses. Guessing in the former may be considered "last-resort" guessing, whereas, after a failure to recall the correct information necessary to determine the correct answer, participants select an alternative without being sure in their choice, but after performing alternative elimination based on information available to them. Conversely, participants who state they guess on the negative-discriminability items are likely performing "first-resort" guessing, where they are selecting an item without necessarily citing anything more complicated than a random guess. This also explains the predictive power of the item discriminability on participant's confidence, where participants who guess on questions immediately viewing it as a deliberate selection, though without a definitive explanation, while

those who guess as after being unable to recall view it as a failure to retrieve information, while test-wiseness may have aided their odds of selecting the right answers. This also matches the gamma correlation between accuracy and discriminability, which showed a significant difference in participants' accuracy on this characteristic.

Hierarchical clustering of the items based on the participants' reported source judgments for the multiple-choice task, along with the rate of accuracy for the items, reveals some division which may relate to the overall difficulty of the items or the efficacy of the first/last-resort strategies that participants used (Figure 10). Given the relatively small amount of variance accounted for, however, there are likely other factors responsible as well.

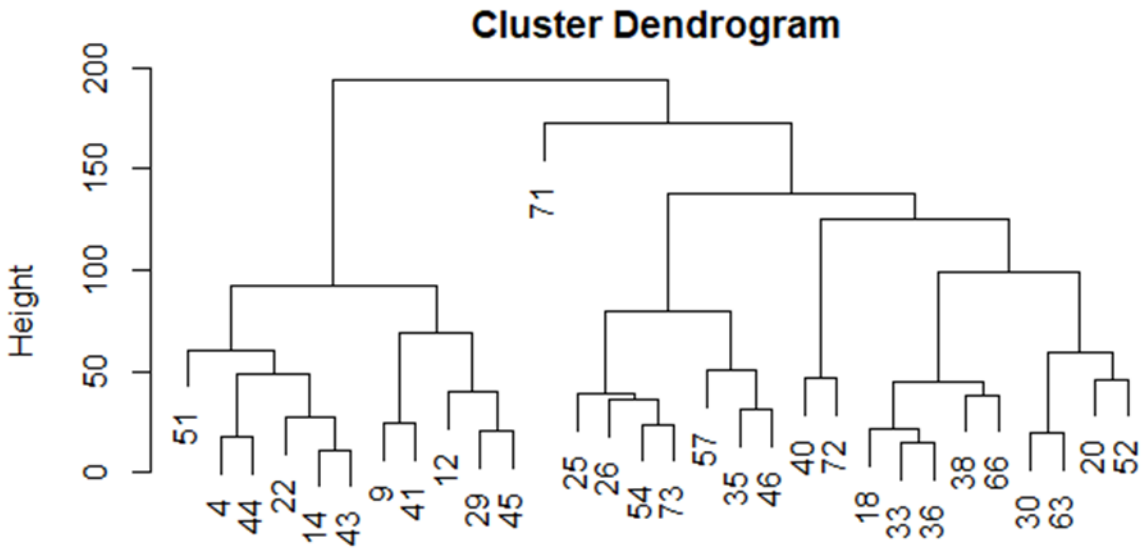

Figure 10. *Hierarchical Clustering of Items*

Hierarchical clustering of items' rates of accuracy and source judgments used. Some division exists, pointing towards specific item characteristics which separate them.

One explanation is the inability of participants to differentiate between a random guess and the use of test-wiseness in making source judgments. The wording of the test-wiseness source ("Answer seemed like a good match for the question") may be approaching the participants' method of answering from the wrong direction (e.g., "Answer seemed like a bad match for the question"). Rather than selecting a single alternative and stating that it is correct, and thus selecting this source, participants likely work towards evaluation of all of the alternatives, removing those that they feel are likely to be incorrect. In this sense, participants are working off of an estimation of probability, and would be more likely to say that they guessed through an arbitrary choice, rather than a definitive selection of what they whole-heartedly believe to be true.

When moving to the multiple-choice task, however, the regression showed that none of the factors served as significant predictors of participant performance. Hierarchical analysis of the participants' accuracies and sources used in the multiple-choice task show no systematicity, with no clear division of the clusters of participants (Figure 11). With additional cues and information being presented to participants due to the format, how can this lack of prediction be explained?

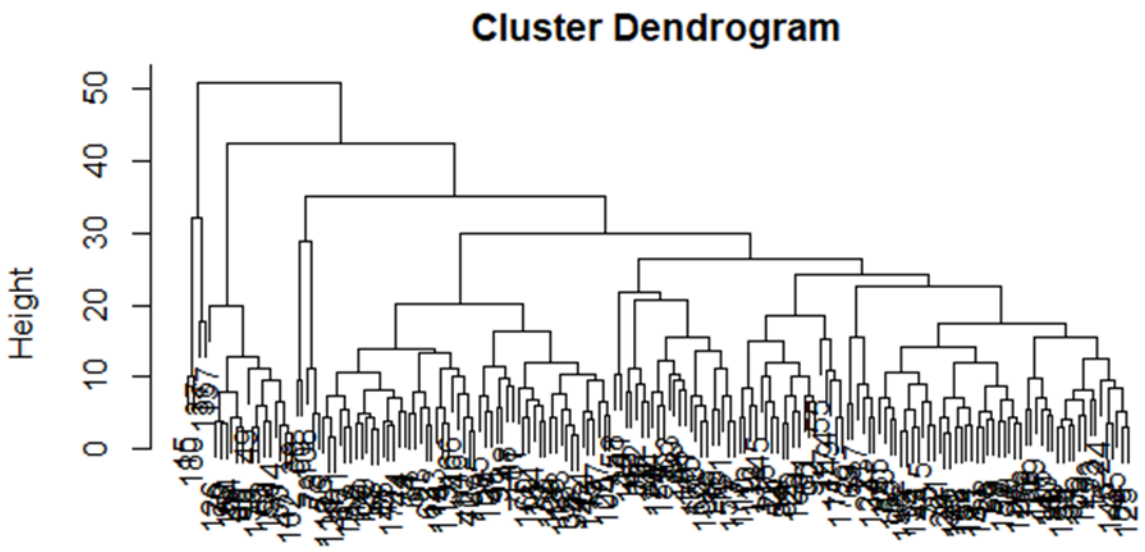

Figure 11. *Hierarchical Clustering of Participants*

Hierarchical clustering of participants' four source attributions and their accuracy on the multiple-choice task. The lack of systematic divisions reflects the lack of predictive power of the four sources and overall scores.

One stated reason is that the differing strategies of the items based on their discriminability leads to no single factor able to predict success across both groups. While formal knowledge may help predict success on some items, its use may be limited by participants' strategies shifting to alternative elimination when prior knowledge is not able to be recalled. Another explanation, describing the lack of predictive power for test-wiseness: upon recognizing the question in the multiple-choice task, participants default to selecting the same answer and source that they used for the previous task, especially if they believe it to be the correct answer, and thus do not use any test-wiseness strategy before answering.

In summary, participants have once again shown an aptitude for answering questions over material that they had presumably never encountered. Even when stating that they had

learnt this material previously, both in formal and informal contexts, their performance showed little difference to those who stated that they merely guessed. These guesses, however, cannot be assumed to be strictly random choices. While self-reports of prior knowledge and test-wiseness were not shown to be significant predictors of test-performance, they must play some role in the ability of participants to perform well on tests, even if participants cannot precisely identify how they have done so. In future research, the key to understanding how these factors truly affect test performance may be further separation of these two sources of information, along with greater identification of which strategy or source participants are using on a per-question basis, possibly through the use of a cognitive architecture to emulate students' test-taking.

REFERENCES

Ashburn, R. (1938). An experiment in the essay-type question. *The Journal of Experimental Education*, *7*(1), 1-3. https://doi.org/10.1080/00220973.1938.11010107

Bacon, D. R. (2003). Assessing learning outcomes: A comparison of multiple-choice and short-answer questions in a marketing context. *Journal of Marketing Education*, *25*(1), 31-36. https://doi.org/10.1177/0273475302250570

Bailey, P. H., Mossey, S., Moroso, S., Cloutier, J. D., & Love, A. (2012). Implications of multiple-choice testing in nursing education. *Nurse Education Today*, *32*(6), e40-e44. https://doi.org/10.1016/j.nedt.2011.09.011

DiBattista, D., & Kurzawa, L. (2011). Examination of the quality of multiple-choice items on classroom tests. *Canadian Journal for the Scholarship of Teaching and Learning*, *2*(2), 4. https://doi.org/10.5206/cjsotl-rcacea.2011.2.4

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. https://doi.org/10.1097/00001888-200210001-00032

Downing, S. M. (2003). Guessing on selected-response examinations. *Medical Education*, *37,* 670-671. https://doi.org/10.1046/j.1365-2923.2003.01585.x

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175-191. https://doi.org/10.3758/BF03193146

Gibb, B. G. (1964). Test-wiseness as secondary cue response. Stanford University.

Hansen, J. D., & Dexter, L. (1997). Quality multiple-choice test questions: Item-writing guidelines and an analysis of auditing testbanks. *Journal of Education for Business*, *73*(2), 94-97. https://doi.org/10.1080/08832329709601623

Hughes, C. A., Salvia, J., & Bott, D. (1991). The Nature and Extent of Test-Wiseness Cues in Seventh—and Tenth-Grade Classroom Tests. *Diagnostique*, *16*(2-3), 153-163. https://doi.org/10.1177/153450849101600310

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2016).lmerTest: Tests in linear mixed effects models. Package version 3.1-1. Retrieved from http://cran.r-project.org/package=lmerTest

Long, J. S. (1997). Regression models for categorical and limited dependent variables (Vol. 7). *Advanced Quantitative Techniques in the Social Sciences*.

Mavis, B. E., Cole, B. L., & Hoppe, R. B. (2001). A survey of student assessment in U.S. medical schools: The balance of breadth versus fidelity. *Teaching and Learning in Medicine, 13,*74-79. https://doi.org/10.1207/S15328015TLM1302_1

McDougall, D. (1997). College Faculty's Use of Objective Tests: State-of-the-Practice versus State-of-the-Art. *Journal of Research and Development in Education*, *30*(3), 183-93.

Millman, J., Bishop, C. H., & Ebel, R. (1965). An analysis of test-wiseness. *Educational and Psychological Measurement*, 25(3), 707-726. https://doi.org/10.1177/001316446502500304

Paxton, M. (2000). A linguistic perspective on multiple choice questioning. *Assessment & Evaluation in Higher Education*, *25*(2), 109-119. https://doi.org/10.1080/713611429

Peirce, J. W., Gray, J. R., Simpson, S., MacAskill, M. R., Höchenberger, R., Sogo, H., Kastman, E., Lindeløv, J. (2019). PsychoPy2: experiments in behavior made easy. *Behavior Research Methods.* https://doi.org/10.3758/s13428-018-01193-y

Roberson, D. (2018). Outsmart the test: Examining the quality of test banks. Unpublished manuscript.

Rogers, W. T., & Bateson, D. J. (1991). The influence of test-wiseness on performance of high school seniors on school leaving examinations. *Applied Measurement in Education*, *4*(2), 159-183. https://doi.org/10.1207/s15324818ame0402_5

Sarnacki, R. E. (1979). An examination of test-wiseness in the cognitive test domain. *Review of Educational Research*, *49*(2), 252-279. https://doi.org/10.3102/00346543049002252

Simkin, M. G., & Kuechler, W. L. (2005). Multiple-choice tests and student understanding: What is the connection?. *Decision Sciences Journal of Innovative Education*, *3*(1), 73-98. https://doi.org/10.1111/j.1540-4609.2005.00053.x

Tarrant, M., Knierim, A., Hayes, S. K., & Ware, J. (2006). The frequency of item writing flaws in multiple-choice questions used in high stakes nursing assessments. *Nurse Education Today*, *26*(8), 662-671. https://doi.org/10.1016/j.nedt.2006.07.006

APPENDIX A

IRB EXEMPTION LETTER

MISSISSIPPI STATE UNIVERSITY

Office of Research Compliance

Institutional Review Board for the Protection of Human Subjects in Research
P.O. Box 6464
56 Morgan Avenue
Mississippi State, MS 39762
P. 662.325.3294

www.orc.msstate.edu

NOTICE OF APPROVAL FOR HUMAN RESEARCH

DATE:	February 07, 2017
TO:	Gary Bradshaw, Ph.D., Psychology
FROM:	Jodi Roberts, HRPP Officer, MSU HRPP
PROTOCOL TITLE:	Can you outsmart the test?
PROTOCOL NUMBER:	IRB-16-697
	Approval Date: February 07, 2017 Expiration Date: December 31, 2019

This letter is your record of the Human Research Protection Program (HRPP) approval of this study as exempt.

On February 07, 2017, the Mississippi State University Human Research Protection Program approved this study as exempt from federal regulations pertaining to the protection of human research participants. The application qualified for exempt review under CFR 46.101(b)(2).

Exempt studies are subject to the ethical principles articulated in the Belmont Report, found at www.hhs.gov/ohrp/regulations-and-policy/belmont-report/#

If you propose to modify your study, you must receive approval from the HRPP prior to implementing any changes. The HRPP may review the exempt status at that time and request an amendment to your application as non-exempt research.

In order to protect the confidentiality of research participants, we encourage you to destroy private information which can be linked to the identities of individuals as soon as it is reasonable to do so.

The MSU IRB approval for this project will expire on December 31, 2019. If you expect your project to continue beyond this date, you must submit an application for renewal of this HRPP approval. HRPP approval must be maintained for the entire term of your project. Please notify the HRPP when your study is complete. Upon notification, we will close our files pertaining to your study.

If you have any questions relating to the protection of human research participants, please contact the HRPP by phone at 325.3994 or email irb@research.msstate.edu. We wish you success in carrying out your research project.

Jodi Roberts

Review Type:	EXEMPT
IRB Number:	IORG0000467