DO TEST ITEMS THAT INDUCE OVERCONFIDENCE MAKE UNSKILLED PERFORMERS UNAWARE?

A dissertation submitted

to Kent State University in partial
fulfillment of the requirements for the
degree of Doctor of Philosophy

by

Marissa K. Hartwig

August, 2013
Dissertation written by
Marissa K. Hartwig
B.A., Ohio Wesleyan University, 2001
M.A., The Ohio State University, 2006
Ph.D., Kent State University, 2013

Approved by

John Dunlosky, Ph.D., Chair, Doctoral Dissertation Committee
Mary Beth Spitznagel, Ph.D., Member, Doctoral Dissertation Committee
Maria S. Zaragoza, Ph.D., Member, Doctoral Dissertation Committee
Timothy Gallagher, Ph.D., Member, Doctoral Dissertation Committee
Christopher Was, Ph.D., Member, Doctoral Dissertation Committee

Accepted by

Maria S. Zaragoza, Chair, Department of Psychology
James L. Blank, Interim Dean, College of Arts and Sciences
# TABLE OF CONTENTS

LIST OF FIGURES........................................................................................................iv

LIST OF TABLES........................................................................................................v

ACKNOWLEDGMENTS..................................................................................................vi

INTRODUCTION............................................................................................................1

- The unskilled-and-unaware pattern: When does it occur?.................................2
- What causes a performer to make an inaccurate global estimate?......................5
- Overview of current studies.................................................................................9

STUDY 1.......................................................................................................................12

- Method..................................................................................................................13
- Results..................................................................................................................14
- Discussion............................................................................................................19

STUDY 2.......................................................................................................................23

- Method..................................................................................................................24
- Results..................................................................................................................26
- Discussion............................................................................................................32

STUDY 3.......................................................................................................................34

- Method..................................................................................................................34
- Results..................................................................................................................36
- Discussion............................................................................................................39

GENERAL DISCUSSION..............................................................................................40

- Conclusion............................................................................................................43

REFERENCES............................................................................................................45

APPENDIX...................................................................................................................49
LIST OF FIGURES

FIGURE 1. Global estimation accuracy, by quartile of performance, for a typical, multiple-choice trivia test (Study 1)........................................................................................................14

FIGURE 2. Mean calibration for a typical, multiple-choice (four-alternative) trivia test (Study 1)...........................................................................................................................................18

FIGURE 3. Mean frequencies of false alarms (FAs) and misses, mean global estimates, and mean global performance, by quartile of performance, for each test group (Study 2)........30

FIGURE 4. Mean calibration of each test group (Study 2)..................................................31

FIGURE 5. Mean frequencies of false alarms (FAs) and misses, mean global estimates, and mean global performance, by quartile of performance, for each test group (Study 3)........38

FIGURE 6. Mean calibration of each test group (Study 3)..................................................39
TABLE 1. Mean frequency of high-confidence judgments, mean of squared deviations between the frequency of high-confidence judgments and global estimates, and correlations between the frequency of high-confidence judgments and global estimates, for each possible boundary (Studies 1-3)........................................................................................................16

TABLE 2. Mean frequency of hits, correct rejections, false alarms, and misses, by quartile of performance (Study 1)........................................................................................................17

TABLE 3. Mean frequency of total responses, correct responses, and incorrect responses, at each confidence level (Study 1)........................................................................................................21

TABLE 4. Expected mean values for performance, overall confidence, and confidence conditionalized on response correctness, for each item set (Study 2)........................................26
ACKNOWLEDGMENTS

My sincere thanks go to my advisor, John Dunlosky, for his guidance, encouragement, and patience during this project and the many projects that preceded it.

I also want to express my appreciation to my family, friends, and colleagues, who have been an important source of support for me. My accomplishments are a reflection of the support I have received from others.
INTRODUCTION

Accurately judging one’s own performance on a test is important for guiding subsequent study decisions. For example, students who believe they performed well on a practice exam – but actually did not – may allocate insufficient study time in preparation for the focal exam. Consistent with this notion, research has demonstrated that participants who more accurately monitor their knowledge make better study decisions (Dunlosky, Hertzog, Kennedy, & Thiede, 2005; Thiede, 1999; Thiede, Anderson, & Therriault, 2003). However, people do not always judge their performance accurately (e.g., Benjamin, Bjork, & Schwartz, 1998; Gigerenzer, Hoffrage, & Kleinbolting, 1991; Lin & Zabrucky, 1998), and individual differences in judgment accuracy have also been observed (e.g., Hartwig, Was, Isaacson, & Dunlosky, 2012; Kruger & Dunning, 1999; Maki, Jonas, & Kallod, 1994). Understanding what factors contribute to accurate vs. inaccurate judgments will benefit efforts to improve those judgments.

A well-known individual difference that often predicts whether performers will make accurate vs. inaccurate judgments is their performance level: Low performers tend to overestimate their overall performance on a test, whereas high performers tend to estimate more accurately or slightly underestimate their performance (Ehrlinger, Johnson, Banner, Dunning, & Kruger 2008; Kruger & Dunning, 1999; Hacker, Bol, Horgan, & Rakow, 2000). The overestimates of low performers have led Kruger and Dunning (1999) to describe these performers as “unskilled and unaware” and “doubly cursed,” i.e., not only do they lack the skills to perform well, but they also lack the metacognitive awareness to recognize their poor performance. In some conditions, however, participants do accurately judge their overall test performance (e.g., Gigerenzer, Hoffrage, & Kleinbolting, 1991), and even unskilled performers...
can be aware of their performance level (Hartwig & Dunlosky, in press). Given the mix of evidence – some showing poor estimation and some showing accurate estimation – an important question arises: What factors determine whether unskilled performers will be aware vs. unaware of their performance level? The answer to this question will be informative both for theory (e.g., How do participants generate their performance estimates?) and for educational application (e.g., In what circumstances are students most in need of feedback to compensate for inaccurate judgments?). Thus, the goals of the present research were to investigate (1) whether participants use their confidence in item responses as a basis for their global estimates, and (2) does high confidence in incorrect responses predict the global unawareness sometimes observed in poor performers? Before describing the present experiments, I briefly review the literature regarding global estimation accuracy of low vs. high performers and highlight reasons to expect that the selection of test items may play a role in producing the unskilled-and-unaware pattern.

The Unskilled-and-Unaware Pattern: When does it occur?

When individuals estimate their overall performance on a test they just completed, low performers frequently overestimate their performance. This tendency has been observed in a variety of tasks, including tests of trivia (Burson, Larrick, & Klayman, 2006), logical reasoning (Kruger & Dunning, 1999), grammar (Kruger & Dunning, 1999), gun safety (Ehrlinger et al., 2008), humor (Kruger & Dunning, 1999), debate (Ehrlinger et al., 2008; Kruger & Dunning, 1999), word prospecting tasks (Burson et al., 2006), classroom exams (Bol & Hacker, 2001; Bol, Hacker, O’Shea, & Allen, 2005; Dunning, Johnson, Ehrlinger, & Kruger, 2003; Hacker, Bol, Horgan, & Rakow, 2000; Hacker, Bol, & Bahbahani, 2008; Nietfeld, Cao, & Osborne, 2005), and other tasks (e.g., Hodges, Regehr, & Martin, 2001). These examples of overestimation by low performers suggest that unskilled performers may indeed be unaware in a variety of situations. Furthermore, the unskilled-and-unaware pattern resonates with the intuition that the skills needed
to perform a task well are often the same skills required to judge the quality of that performance (Kruger & Dunning, 1999).

However, recent studies indicate that unskilled performers are not always inaccurate when estimating their performance. For example, Burson et al. (2006) demonstrated that the level of performer appearing most accurate can be manipulated by altering the difficulty of the task. Specifically, easier tasks made low performers appear unaware (overestimation) and high performers appear more accurate, whereas difficult tasks made high performers appear unaware (underestimation) and low performers appear more accurate. These results suggest that the accuracy of performance estimates may be linked to test characteristics and not produced by individual differences in awareness.

Further evidence that unskilled performers are not always unaware has been provided by Hartwig and Dunlosky (in press). In several studies, low performers (and, in fact, all performers) were consistently accurate in estimating how many test questions they had answered correctly, which suggests that the unskilled—and-unaware pattern may be less robust than previously thought. Furthermore, they demonstrated that the response scale on which participants estimate their performance is consequential: On a variety of tasks (including tests of general trivia, algebra story problems, and introductory psychology materials), participants at all levels of performance were highly accurate when estimating the number of questions answered correctly, but those very same low and high performers made overestimates and underestimates, respectively, when estimating their performance on a scale of percentile rank. The authors concluded that, in contrast to number-correct estimates, percentile-rank estimates require that participants estimate the quality of others’ performance, in addition to their own performance, to produce a final estimate of relative standing. Given that the very same performers in those studies made accurate estimates of their absolute performance (number correct) but inaccurate estimates of relative
performance (percentile rank), caution is required in interpreting “unawareness” in previous studies that have used only percentile rank scales to demonstrate the unskilled-and-unaware pattern. Specifically, the inaccuracy observed in percentile rank estimates may arise from difficulty in estimating normative performance and does not necessitate unawareness of one’s own performance.

Nonetheless, the overestimates of unskilled performers are not strictly limited to percentile rank estimates. Absolute estimates (such as number-correct or percent-correct estimates) have been included in some studies of the unskilled-and-unaware pattern, and results are mixed: Low performers overestimate their performance in some studies (e.g., Ehrlinger et al., 2008) but not in others (e.g., Hartwig&Dunlosky, under review). These studies were similar in procedure, but have differed in their criterion tests of performance (e.g., tests of logical reasoning yielded the unskilled-and-unaware pattern, whereas tests of paired associates did not), which might suggest that differences in test content are responsible for the mixed results. However, studies using similar tests have also produced mixed results. For example, tests of University-of-Chicago trivia produced the unskilled-and-unaware pattern, whereas tests of general-knowledge trivia did not. This unexpected discrepancy suggests that the critical differences in content may pertain to the specific items included in a test.

In summary, unskilled performers may overestimate their performance for some tests – but not all tests. With respect to estimates of absolute performance (which is the focus of the present research), what tests will vs. will not produce the unskilled-and-unaware pattern remains unclear. To help formulate predictions about when the pattern will occur (i.e., for what tests or sets of items), the next section will examine several theories that have been proposed to explain why inaccurate global estimates occur.
What causes a performer to make an inaccurate global estimate?

Kruger and Dunning (1999) proposed that low performers’ inaccurate global estimates are caused by a deficit in metacognitive skills. They operationalized metacognitive skills as “(a) the ability to distinguish what one has answered correctly from what one has answered incorrectly and (b) the ability to recognize competence in others.” They provided preliminary evidence for this metacognitive deficit hypothesis by demonstrating that participants’ skill in identifying correct vs. incorrect answers partially mediated the link between performance level and global estimation accuracy. Also, they showed that low (vs. high) performers were less accurate in grading the test performance of their peers and gained less insight about their own level of performance from doing so. Thus, they concluded that low performers overestimate their performance due to poor metacognitive skill. However, Hartwig and Dunlosky (under review) experimentally tested an instantiation of this hypothesis and found that poor item discrimination was not sufficient to guarantee poor global estimation accuracy. Specifically, they manipulated participants’ ability to discriminate correct from incorrect responses by varying test content (paired associates tests vs. trivia tests) and response formats (recall tests vs. recognition tests) between groups. Although these manipulations produced differences in discrimination accuracy, conditions that reduced discrimination accuracy did not produce inaccurate global estimates.

Nonetheless, the assumption that item confidence judgments are somehow a basis for global estimation has some support (e.g., Dunlosky & Hertzog, 2000), so Hartwig and Dunlosky (under review) proposed an alternative hypothesis: The item-frequency hypothesis assumes that when people are asked to estimate how many test items they answered correctly, they compute the frequency of times that item judgments produced high confidence. For example, a participant may compute that they had high confidence (e.g., confidence greater than 5 on a scale of 0-10) for approximately 25 of the 40 test questions, thus leading to a global estimate of 25 items correct.
This simple hypothesis has implications for the relationship between item-judgment accuracy and global estimation accuracy: Namely, item responses that yield high confidence when actually incorrect (aka false alarms, or FAs) will contribute to global overestimates, whereas item responses that yield low confidence when actually correct (aka misses) will contribute to global underestimates. To better understand how this pattern would emerge, imagine that a participant had high confidence in 25 (out of 40) items (and low confidence in the remaining 15 items), as in the example above, producing a global estimate of 25. Of those high-confidence items, perhaps only 15 were actually correct; the other 10 were incorrect (FAs) and would contribute to global overestimation. However, overestimation due to FAs can be offset by underestimation due to misses: For example, perhaps 4 of the low-confidence items were unexpectedly correct. The actual total correct – i.e., the sum of correct items, regardless of confidence – would be 19 (15 high-confidence correct items plus 4 low-confidence correct items). Thus, the participants’ global estimate of 25 would overestimate their actual performance by 6 items. This magnitude of overestimation (6 items) corresponds to the frequency of FAs (10) minus the frequency of misses (4). In contrast, if misses had outnumbered FAs, then underestimation would be expected; and if FAs and misses occurred with similar frequencies, then accurate global estimation would be expected.

Thus, a corollary to the item-frequency hypothesis is that the frequency of FAs and the frequency of misses will approximately determine global estimation accuracy. Hartwig and Dunlosky (under review) provided preliminary support for the hypothesis by demonstrating that FAs and misses were similar in frequency (offsetting one another) in several experiments which had all produced accurate global estimates. However, they did not observe dissimilar frequencies of FAs and misses, so they could not evaluate the most critical prediction from the item-frequency hypothesis, namely, that global estimation accuracy will vary with the ratio of FAs and
misses. In particular, global overestimates are predicted when FAs exceed misses, whereas
global underestimates are predicted when misses exceed FAs. Thus, a major goal of the current
research was to test these expectations by manipulating sets of items to contain many FA-
inducing items, many miss-inducing items, both, or neither. Importantly, if expectations are met,
the item-frequency hypothesis offers specific predictions about how item judgments and global
estimates are linked, as well as a plausible reason for why some tests have produced the
unskilled-and-unaware pattern and others have not. Namely, tests that contain many “tricky”
items – i.e. items likely to produce FAs – will promote global overestimates. And because low
(vs. high) performers, by definition, answer more items incorrectly, low performers may be at
greater risk for FAs (and, thus, global overestimation) on these tricky tests.

Hartwig and Dunlosky (under review) are not the first to propose that the selection of test
items can influence the accuracy of participants’ judgments. Gigerenzer, Hoffrage, and
Kleinbolting (1991) developed a theory of probabilistic mental models (PMM) to explain the
overconfidence typically observed in calibration curves derived from participants’ item
confidence judgments. These calibration curves show that for items a participant rates with $X\%$
confidence, the participant typically gets less than $X\%$ of those items correct, indicating
miscalibration (specifically, overconfidence). PMM theory explains this item overconfidence as
follows: When participants are unable to retrieve an item response directly from memory or to
use logical operations to obtain a response with complete certainty, participants instead select
their answer based on probabilistic information encoded from previous interaction with their
environment. For example, in a two-alternative general knowledge question, participants may
need to judge which of two American cities is larger in population: San Jose or Baltimore. If
participants cannot retrieve population information directly from memory to produce a certain
response, then they will instead rely on cues which they have previously learned are correlated
with population. Thus, the participant searches for a cue that will (a) discriminate between the two cities and (b) allow an inference regarding relative population. A cue – such as whether a city has a professional sports team – may have good ecological validity for predicting city population; and, according to PMM theory, the validity of such cues become internalized through interaction with one’s environment. Relying on these cues will allow a participant to select a response, and the internalized cue validity will determine the confidence associated with that response (i.e., cues perceived to have high validity will result in high confidence in the selected response). In the example above, because Baltimore has several well-known professional teams and San Jose does not, the participant may select Baltimore as the larger city, with reasonably high confidence due to the perceived validity of the cue. According to PMM theory, item overconfidence typically occurs not because participants overestimate the validity of the cues in the real environment, but because trivia test items are usually non-randomly selected and contain many “tricky” items that violate expectations derived from cues (e.g., San Jose actually exceeds Baltimore in population, inconsistent with the sports-team cue). In other words, the cities compared on a test of city populations are often not a representative sample of all possible city comparisons in the reference set; rather, they are selected to be tricky. If, instead, the test items were representatively sampled from the environment, then accurate item calibration would be expected, because the perceived cue validities (and their associated confidence judgments) would be appropriate for the task. In support of these predictions, Gigerenzer et al. (1991) demonstrated that representative (non-tricky) item sets produced good item calibration, whereas selected item sets produced the overconfidence that is often observed.

The above description of PMM theory pertains to item confidence judgments, but the theory was also applied to global estimates. Specifically, Gigerenzer et al. (1991) suggested that when participants make global estimates, they again rely on information acquired through past
interaction with their environment. In this case, however, they rely on their past experience with all similar trivia tests (e.g., How do I typically score on tests like this?) rather than focusing on test items. Because trivia tests tend to contain tricky items, participants’ past experience with trivia tests reflects this tendency and leads participants to make accurate global estimates. Importantly, these accurate global estimates (for tricky tests) co-occur with miscalibrated (overconfident) item judgments, a phenomenon referred to as the confidence-frequency effect. Gigerenzer et al. (1991) also demonstrated that if the test instead included a representative sample of questions (non-tricky test), participants now underestimated their global performance and their item judgments were well calibrated. Extrapolating further, if the test were exceptionally tricky – even more than usual – then participants should overestimate their global performance and item calibration should show exaggerated overconfidence.

Thus, both the item-frequency hypothesis (Hartwig and Dunlosky, under review) and Gigerenzer et al.’s PMM (1991) proposed that the researcher’s selection of test items can predictably influence global estimation accuracy. However, the two perspectives have numerous differences. Several important differences are described below and provide direction for the present studies.

**Overview of Current Studies**

A primary difference between the two perspectives pertains to their goals. Gigerenzer et al. (1991) proposed a detailed theory of item calibration and global estimation, but they focused on group averages rather than differences between low and high performers. In contrast, Hartwig and Dunlosky (under review) aimed to understand why individuals sometimes differ in their global estimation accuracy and identify conditions that may produce the unskilled-and-unaware phenomenon. The current studies share this latter goal.
Another important difference pertains to the hypothesized basis of global estimates. Gigerenzer et al.’s (1991) PMM theory assumes that item confidence and global estimates are largely unconnected – i.e., item confidence results from the perceived validity of the utilized cues, whereas global estimates result from estimating past performance on similar tests. How item confidence in a current task might influence a current global estimate was not explained. In contrast, Hartwig and Dunlosky (under review) hypothesized a direct connection between item confidence and global estimates: The item-frequency hypothesis proposes that high-confidence items in the current task are summed to generate a global estimate. Furthermore, a corollary of this hypothesis is that the frequencies of FAs and misses on a test should approximately determine global estimation accuracy. The present studies were designed to evaluate the item-frequency hypothesis by testing the following predictions derived from it: (1) the frequency of high-confidence responses will correspond to participants’ global estimates, and (2) the balance of false alarms vs. misses will predict the accuracy of the global estimates observed. To foreshadow, in Study 1, a typical trivia test was administered, and both predictions were met. In Studies 2 and 3, several different trivia tests were created to produce varying frequencies of FAs and misses. Tests yielding more FAs than misses were expected to produce global overestimates; tests yielding more misses than FAs were expected to produce global underestimates; and tests yielding a balance of FAs and misses were expected to produce accurate global estimation.

A final difference between the two perspectives pertains to their analysis of item judgments. Like others who have investigated calibration, Gigerenzer et al. (1991) evaluated item-judgment accuracy via calibration curves, which plot the level of judged confidence (x-axis) against the level of performance (y-axis). For typical trivia tests, these calibration curves consistently show substantial overconfidence (and little or no underconfidence). In contrast, Hartwig and Dunlosky (under review) did not use calibration curves to evaluate item accuracy,
but instead computed the frequencies of FAs (overconfidence) and misses (underconfidence), which were observed to be approximately balanced in a typical trivia test. These two observations about item judgments (i.e., predominant overconfidence vs. a balance of overconfidence and underconfidence) seem discrepant. Thus, a secondary goal of the present studies was to investigate this apparent discrepancy. To do so, item accuracy was evaluated in both ways – calibration curves and frequency analyses – in the present studies. By comparing these analyses, we not only resolve the apparent discrepancy, but we also gain insight into how the item-frequency hypothesis might be improved.
STUDY 1

The primary goal of Study 1 was to evaluate the item-frequency hypothesis for a typical trivia test. The trivia test was a four-alternative multiple-choice test previously used by Hartwig and Dunlosky (under review), and, thus, participants’ global estimates were expected to be accurate (replicating their results). To evaluate the item-frequency hypothesis, I first tested whether the frequency of high-confidence responses corresponded to participants’ global estimates, which is necessary for the hypothesis to be supported. Second, I evaluated whether the frequencies of false alarms vs. misses would predict the accuracy of global estimates (a corollary of the hypothesis). Because global estimates were accurate on this test, I expected FAs and misses to be approximately balanced (offsetting one another). These expectations applied to the group of test-takers as a whole, as well as to each quartile of performers.

In addition to the frequency analysis of FAs and misses just described, a calibration-curve analysis (alaGigerenzer et al., 1991) was conducted for comparison. The calibration literature indicates that item overconfidence is typically dominant when global estimates are accurate (Gigerenzer et al., 1991), an observation that seems inconsistent with Hartwig and Dunlosky’s (under review) observation of balanced overconfidence (FAs) and underconfidence (misses). A possible reason for this discrepancy is that different trivia tests were used in the respective studies. Thus, one possible outcome is that the current trivia test will not produce the typical pattern of item miscalibration (i.e., overconfidence) and instead may produce a combination of overconfidence and underconfidence, analogous to the balance of FAs and misses. Another possibility, however, is that calibration curves will indeed show the typical pattern of
overconfidence, even while FAs and misses are approximately balanced. If so, we will consider how these seemingly discrepant observations can be reconciled.

Method

Participants and Design

Forty-four undergraduate students participated for course credit. They were part of a participant pool, which consisted mainly of students enrolled in introductory psychology or other low-level psychology courses.

Materials

Materials were 40 general knowledge questions (e.g., What kind of metal is associated with a 50th wedding anniversary?) from Nelson and Narens’ (1980) norms and were the same items used by Hartwig and Dunlosky (under review). Each question could be answered with a one-word response and was presented along with four response options in random order. One option was the correct answer (gold); the other three options were plausible but incorrect responses (silver, platinum, bronze). The 40 questions were selected to represent a range of difficulty from very easy to difficult. The tendency for an item to produce either FAs or misses was not used as a criterion for item selection in Study 1.

Procedure

Participants worked individually at computers. They first read instructions on the computer screen that described what they would be asked to do.

Test phase. After reading instructions, 40 trivia questions were presented, one at a time, in random order. Each question was presented with four response options, and participants were required to select one of the four options. After responding, participants rated their confidence in the selected response (from 0 = not confident at all to 10 = completely confident that their response was correct). Testing was self-paced.
Global estimates. At the conclusion of the test phase, participants were asked to make global estimates about their performance on the test as a whole. First, they estimated the number of items (out of the 40) that they had answered correctly. Second, they estimated the percent of items that they answered correctly on the test (from 0 to 100%). Results from these two global estimates provided nearly identical results; thus, only estimates of number correct will be presented.

Results

Accuracy of global estimates

First, I verified that global estimates were highly accurate for this trivia test, replicating Hartwig and Dunlosky (under review). Indeed, participants’ estimated performance (\(Mean = 23.3\) items) did not differ from their actual performance (\(Mean = 22.8\) items), \(t(43) = .46, p = .65\). Furthermore, accuracy did not differ by quartile of performance, \(F(3,40) = 1.35, p = .27\), (Figure 1). Thus, even low performers estimated their overall performance accurately.

![Fig. 1. Global estimation accuracy, by quartile of performance, for a typical, multiple-choice trivia test (Study 1).](image)
Computing frequencies of false alarms and misses

Next, I analyzed item-judgment accuracy by computing the frequency of false alarms (FAs) and misses for each individual. In previous research, Hartwig and Dunlosky (under review) computed FAs and misses as follows: Incorrect responses given confidence ratings of 6 or above (considered high confidence on a 0-10 scale) were counted as FAs, and correct responses given confidence ratings of 5 or below (low confidence) were counted as misses. Although the mid-point of the confidence scale may be a plausible boundary for separating responses that were high-confidence (perceived to be correct) vs. low-confidence (perceived to be incorrect), this classification may not match how the scale was actually used by some participants. Thus, the present study determined the boundary empirically. First, I computed the frequency of high-confidence judgments that would result from each possible boundary. The average of these frequencies, their squared deviation from participants’ global estimates, and their correlation with participants’ global estimates are presented in Table 1. I then selected the boundary that produced the best correspondence between the frequency of high-confidence item judgments and participants’ global estimates. The selected boundary (i.e., values ≥6 denote high confidence) is actually the same boundary used by Hartwig and Dunlosky (under review). This boundary was used in all subsequent analyses.

Evaluating the item-frequency hypothesis

The item-frequency hypothesis posits that participants use the frequency of their high-confidence item judgments as an estimate of their global performance. Consistent with this hypothesis, the association between the participants’ global estimates and the frequency of high-confidence judgments was strong (r = .83).

---

1 By selecting the boundary in this way, we are giving the item-frequency hypothesis its best chance at receiving support, though support is still not guaranteed. Studies 2 and 3 will further address the question of which boundary is appropriate by asking participants to explicitly dichotomize their confidence judgments into high- vs. low-confidence, in addition to rating their confidence on a 0-10 scale.
Table 1. Mean frequency of high-confidence judgments, mean of squared deviations between the frequency of high-confidence judgments and global estimates, and correlations between the frequency of high-confidence judgments and global estimates, for each possible boundary (Studies 1-3).

<table>
<thead>
<tr>
<th>Study</th>
<th>Possible boundaries for defining high confidence (0-10 scale)</th>
<th>( \geq 1 )</th>
<th>( \geq 2 )</th>
<th>( \geq 3 )</th>
<th>( \geq 4 )</th>
<th>( \geq 5 )</th>
<th>( \geq 6 )</th>
<th>( \geq 7 )</th>
<th>( \geq 8 )</th>
<th>( \geq 9 )</th>
<th>( \geq 10 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study 1</td>
<td>Mean frequency of high-confidence judgments (should approximate the mean global estimate made by participants: 23.3)</td>
<td>36.5</td>
<td>34.5</td>
<td>32.1</td>
<td>29.6</td>
<td>27.3</td>
<td>24.8</td>
<td>22.0</td>
<td>19.0</td>
<td>15.7</td>
<td>11.9</td>
</tr>
<tr>
<td></td>
<td>Mean of squared deviations between frequency of high-confidence judgments &amp; global estimates (minimal values desired)</td>
<td>223.4</td>
<td>168.3</td>
<td>111.0</td>
<td>66.8</td>
<td>42.7</td>
<td>27.5</td>
<td>29.5</td>
<td>49.0</td>
<td>95.9</td>
<td>176.3</td>
</tr>
<tr>
<td></td>
<td>Correlation between frequency of high-confidence judgments &amp; global estimates (maximum correlation desired)</td>
<td>.612</td>
<td>.659</td>
<td>.757</td>
<td>.813</td>
<td>.818</td>
<td>.830</td>
<td>.817</td>
<td>.789</td>
<td>.725</td>
<td>.637</td>
</tr>
<tr>
<td>Study 2</td>
<td>Mean frequency of high-confidence judgments (should approximate the mean global estimate made by participants: 24.5)</td>
<td>37.6</td>
<td>36.4</td>
<td>34.6</td>
<td>32.5</td>
<td>30.2</td>
<td>26.0</td>
<td>23.0</td>
<td>19.6</td>
<td>15.8</td>
<td>12.8</td>
</tr>
<tr>
<td></td>
<td>Mean of squared deviations between frequency of high-confidence judgments &amp; global estimates (minimal values desired)</td>
<td>223.2</td>
<td>186.5</td>
<td>144.2</td>
<td>107.1</td>
<td>73.0</td>
<td>37.5</td>
<td>39.2</td>
<td>64.6</td>
<td>126.2</td>
<td>198.8</td>
</tr>
<tr>
<td></td>
<td>Correlation between frequency of high-confidence judgments &amp; global estimates (maximum correlation desired)</td>
<td>.380</td>
<td>.492</td>
<td>.563</td>
<td>.588</td>
<td>.634</td>
<td>.694</td>
<td>.685</td>
<td>.657</td>
<td>.577</td>
<td>.481</td>
</tr>
<tr>
<td>Study 3</td>
<td>Mean frequency of high-confidence judgments (should approximate the mean global estimate made by participants: 9.6)</td>
<td>18.0</td>
<td>17.0</td>
<td>15.8</td>
<td>13.9</td>
<td>12.6</td>
<td>10.4</td>
<td>8.9</td>
<td>7.4</td>
<td>5.6</td>
<td>4.1</td>
</tr>
<tr>
<td></td>
<td>Mean of squared deviations between frequency of high-confidence judgments &amp; global estimates (minimal values desired)</td>
<td>87.5</td>
<td>71.8</td>
<td>52.7</td>
<td>33.1</td>
<td>23.7</td>
<td>14.5</td>
<td>15.1</td>
<td>18.4</td>
<td>30.0</td>
<td>42.8</td>
</tr>
<tr>
<td></td>
<td>Correlation between frequency of high-confidence judgments &amp; global estimates (maximum correlation desired)</td>
<td>.331</td>
<td>.368</td>
<td>.496</td>
<td>.557</td>
<td>.561</td>
<td>.618</td>
<td>.600</td>
<td>.615</td>
<td>.589</td>
<td>.576</td>
</tr>
</tbody>
</table>
A corollary of this hypothesis is that the balance of the frequencies of FAs and misses should predict the accuracy of participants’ global estimates. The mean number of FAs and misses for each quartile of performers is shown in the bottom two rows of Table 2. For quartiles 1-3, the frequencies of FAs did not differ from the frequencies of misses, (all ps > .05), which is consistent with the item-frequency hypothesis because balanced frequencies of FAs and misses would be expected to produce global estimates that are accurate. In quartile 4, however, the frequency of FAs (Mean = 6.0) slightly exceeded the frequency of misses (Mean = 3.1), \( t(10) = 3.17, p = .01 \), an unexpected result that is inconsistent with previous research.

Table 2. Mean frequency of hits, correct rejections, false alarms, and misses, by quartile of performance (Study 1).

<table>
<thead>
<tr>
<th></th>
<th>Q1 (n=10)</th>
<th>Q2 (n=11)</th>
<th>Q3 (n=12)</th>
<th>Q4 (n=11)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Hit</strong> (high confidence, correct response)</td>
<td>8.5</td>
<td>17.6</td>
<td>19.8</td>
<td>26.5</td>
</tr>
<tr>
<td><strong>CR</strong> (low confidence, incorrect response)</td>
<td>18.8</td>
<td>11.4</td>
<td>9.1</td>
<td>4.5</td>
</tr>
<tr>
<td><strong>FA</strong> (high confidence, incorrect response)</td>
<td>5.8</td>
<td>7.7</td>
<td>6.3</td>
<td>6.0</td>
</tr>
<tr>
<td><strong>Miss</strong> (low confidence, correct response)</td>
<td>6.9</td>
<td>3.4</td>
<td>4.8</td>
<td>3.1</td>
</tr>
</tbody>
</table>

Nevertheless, the difference between the two frequencies was small (less than a 3-item difference) and may not represent a sufficiently large imbalance to affect global accuracy on a 40-item test. In fact, the difference in quartile 4 (Mean = 2.9) was smaller than the difference in quartile 2 (Mean = 4.3, which was not significant). The significant result in quartile 4 may reflect less variability in the frequencies of FAs among high performers (SD = 2.2 in Q4 vs. 4.8 in Q2), whose high scores are closer to the maximum score and thereby limit the number of FAs that can possibly occur. Thus, if we assume larger imbalances are required to influence global accuracy, the item-frequency hypothesis could still account for these observations. In Studies 2 and 3, to
provide a stronger test of the hypothesis, I attempted to create larger discrepancies between the frequencies of FAs and misses.

*Calibration curve*

Item-judgment accuracy was also evaluated by constructing a calibration curve for each individual and then averaging across all participants. In a calibration curve, each possible confidence level (0-10, shown on the x-axis) is plotted against the percent of items answered correctly at that confidence level (y-axis). Accuracy is typically defined as a diagonal line with a slope of +1 (with an intercept of 0), such that judgments are considered optimal when, for example, a rating of 70 out of 100 (or 7 out of 10 on the current scale) is associated with getting 70% of items correct. Figure 2 displays an average of the individual calibration curves, as well as the accuracy line just described (solid diagonal). However, because the test was a 4-alternative multiple-choice test, even a randomly-selected guess would be correct approximately 1 out of 4 times. Thus, the percent of questions answered correctly at the lowest confidence level should not fall below 25%, and a revised accuracy line (dotted diagonal) takes correct guessing into account.

![Figure 2. Mean calibration for a typical, multiple-choice (four-alternative) trivia test (Study 1).](image-url)
Based on this analysis, I concluded that overconfidence was substantial: When participants judged their responses with a particular level of confidence, they answered fewer items correctly than expected for that confidence level. No underconfidence was evident (based on the corrected accuracy line). Thus, the calibration analysis indicated that participants’ item judgments were overconfident (consistent with Gigerenzer et al. 1991), whereas the frequency analysis indicated that instances of FAs (overconfidence) and misses (underconfidence) were approximately balanced (consistent with Hartwig and Dunlosky, under review). The reason for the apparent discrepancy of these outcomes will be clarified in the Discussion section below.

*Item discrimination*

For consistency with previous research, I also evaluated item discrimination, or the degree to which higher confidence ratings are associated with correct responses and lower ratings are associated with incorrect responses. A gamma correlation (see Nelson, 1984) was computed for each individual, where values close to +1 indicate good item discrimination. As in previous research (Hartwig and Dunlosky, under review), mean item discrimination was above chance (*Mean gamma* = .58, *SE* = .03) and differed by quartile of performance, *F*(3,40) = 5.68, *p* = .002, such that lower performers demonstrated poorer discrimination than higher performers (*Q1 gamma* = .37, *SE* = .08; *Q4 gamma* = .65, *SE* = .04).

*Discussion*

The results of Study 1 were largely consistent with the item-frequency hypothesis. Using a boundary at the middle of the scale to separate low and high confidence, the frequency of high-confidence judgments was strongly correlated with participants’ global estimates. Furthermore, the frequencies of FAs and misses were approximately balanced and co-occurred with accurate global estimates, which is also consistent with the hypothesis. Thus, Study 1 provides preliminary support for the item-frequency hypothesis. In Studies 2 and 3, I perform more
compelling tests of the hypothesis by experimentally manipulating the frequencies of FAs and misses and then observing the effect on global estimation accuracy.

Although an approximate balance of FAs and misses was observed in this study, the calibration curve demonstrated overconfidence only. Both results are consistent with previous research, but seem discrepant from one another. To reconcile these results, I consider next how item judgment accuracy is defined in each type of analysis.

In typical calibration analyses, participants’ item judgments are considered accurate if the confidence level corresponds to the percent of items answered correctly. For example, items rated at a confidence level of 80 out of 100 (or 8 out of 10) should be answered correctly 80% of the time. But is this definition of accuracy appropriate? Some researchers have questioned this interpretation of confidence ratings and have provided some evidence that participants may not spontaneously use numerical response scales in such a sophisticated manner (e.g., Ronis & Yates, 1987; Zimmer, 1983). Other evidence, however, suggests that this interpretation is reasonable (e.g., Gigerenzer et al., 1991). Most important, note that calibration curves (1) focus on percentages (rather than frequencies) of correct responding at various confidence levels, and (2) define judgment accuracy as a match between the confidence level and the percentage of correct answers that should be earned at that confidence level (corrected for guessing, in the present case).

In contrast, the analysis of FAs and misses focuses on frequencies rather than percentages. The analysis does not require each confidence level to be evaluated separately (as done in calibration curves), but only requires a distinction be made between low and high confidence. For accuracy to occur, any item rated with confidence above 5 (the midpoint) is expected to be correct, and any item rated with confidence of 5 or below is expected to be incorrect. For comparison to the calibration analysis, imagine how perfect accuracy for the
frequency analysis would appear if drawn in Figure 2: After converting frequencies to percentages, perfect accuracy would not be represented by a diagonal, but rather would be represented by a step function in which a single step up (from 0% to 100%) occurred between the 5- and 6-level of confidence (corresponding to the midpoint boundary). Any deviation from 100% at a confidence of 6 or above would indicate the presence of FAs, and any deviation from 0% at a confidence of 5 or below would indicate the presence of misses. Thus, in the present data, the observed percentages in Figure 2 indicate that both FAs and misses do occur. The exact frequencies of FAs and misses are not evident in the figure because only proportions of items, and not absolute frequencies, are conveyed. Furthermore, the various confidence levels are used with different frequencies. To further illustrate the similarities and differences between the two analyses, Table 3 displays a summary of the underlying data. The calibration curve reflects the number of correct responses divided by the total number of responses at each confidence level.

Table 3. Mean frequency of total responses, correct responses, and incorrect responses, at each confidence level (Study 1).

<table>
<thead>
<tr>
<th>Confidence level</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total responses</td>
<td>3.5</td>
<td>2.1</td>
<td>2.3</td>
<td>2.5</td>
<td>2.4</td>
<td>2.5</td>
<td>2.8</td>
<td>3.0</td>
<td>3.3</td>
<td>3.8</td>
<td>11.9</td>
</tr>
<tr>
<td>Correct responses</td>
<td><strong>0.9</strong></td>
<td><strong>0.4</strong></td>
<td><strong>0.7</strong></td>
<td><strong>0.8</strong></td>
<td><strong>0.8</strong></td>
<td><strong>0.9</strong></td>
<td>1.4</td>
<td>1.7</td>
<td>2.1</td>
<td>2.8</td>
<td>10.4</td>
</tr>
<tr>
<td>Incorrect responses</td>
<td>2.6</td>
<td>1.6</td>
<td>1.7</td>
<td>1.7</td>
<td>1.5</td>
<td>1.6</td>
<td><strong>1.3</strong></td>
<td><strong>1.3</strong></td>
<td><strong>1.3</strong></td>
<td><strong>1.1</strong></td>
<td><strong>1.5</strong></td>
</tr>
</tbody>
</table>

At most confidence levels, overconfidence is apparent (in comparison to the corrected accuracy line). In contrast, the frequency analysis involves summing the bold-faced values in the lower right portion of the table to compute false alarms and summing the bold-faced values in the

---

However, the percentages computed from Table 3 will not perfectly match the percentages plotted in the calibration curve, because the calibration curve was calculated by averaging across individuals’ percentages, rather than by calculating percentages from average frequencies.
middle left portion of the table to compute misses. Here, FAs and misses can be seen to occur with similar frequency. Thus, the two analyses are not actually contradictory, but rather they communicate different features of the data.

A limitation of both analyses described here is that the participants’ interpretation of the confidence scale is not known with certainty. In the calibration analysis, a discrepancy between confidence level (e.g., 8, or 80%) and performance level (70%) may not convey true overconfidence if the participant is using the scale as an ordinal scale, rather than an interval scale that assumes values correspond to proportion correct. A limitation in the frequency analysis was that a participants’ true boundary between high and low confidence judgments (which affects our computation of the frequencies of FAs and misses) is not known and may even differ by participant. Thus, Studies 2 and 3 supplement the 0-10 confidence judgment with another judgment. Specifically, after making a 0-10 confidence judgment for a given item, participants were asked to also indicate whether they believed their response was more likely correct or more likely incorrect. By making explicit the judgment of perceived correctness, we will be able to more definitively evaluate the presence of FAs and misses.
STUDY 2

The observations in Study 1 were generally consistent with the item-frequency hypothesis proposed by Hartwig and Dunlosky (under review), but the hypothesis remained experimentally untested. Thus, the primary goal of Study 2 was to manipulate frequencies of FAs and misses to observe the effect on global estimates. If the tendency of some items to produce either FAs or misses does influence the accuracy of global estimates, then the manipulations should produce corresponding changes in global estimation accuracy. To test this possibility, I created several tests comprising different sets of items, including sets that would produce greater frequencies of FAs (vs. misses) or greater frequencies of misses (vs. FAs). According to the hypothesis, tests in which FAs exceed misses should produce global overestimates; tests in which misses exceed FAs should produce global underestimates; and tests containing similar frequencies of each should produce global accuracy.

To create these sets of items, I gathered item statistics from several previous studies that had used multiple-choice trivia items (including Study 1). These statistics included mean confidence for incorrect responses and mean confidence for correct responses, which allowed for selection of items that would produce item-overconfidence (FAs) and item-underconfidence (misses), respectively. Will these sets produce corresponding over- and under-estimates of global performance? Also, will these sets show differing effects for low (vs. high) performers? Given that low (vs. high) performers produce more incorrect responses, they have more opportunity for FAs and may be more affected by manipulations designed to produce overconfidence in wrong answers. Similarly, high performers (who produce more correct responses) have more opportunity for misses and may be more affected by manipulations
designed to produce underconfidence in correct answers. If these manipulations cause high and low performers to differ in their global estimation accuracy, item selection could provide a plausible explanation for why the unskilled-and-unaware pattern has occurred for some trivia tests but not others.

Method

Participants and Design

Participants were 201 undergraduate students who participated for course credit via the participant pool. The study was conducted in the laboratory. Four different sets of items were constructed to serve as criterion tests (described in the Materials section below), and participants were randomly assigned to receive one of the four tests: (1) the FA-items test, (2) the miss-items test, (3) the accuracy-items test, or (4) the half-FA/half-miss-items test.

Materials

Each of the four tests was composed of 40 general knowledge questions drawn from Nelson and Narens’ (1980) norms. The selection of items for each test was based on an item analysis (described below) of 80 items used in previous studies. These items represented a range of difficulty. Each item could be answered with a one-word response and was presented along with four plausible response options in random order.

Analysis of items from previous studies. For each trivia item used in prior studies, I computed the proportion of times it was answered correctly vs. incorrectly, the mean confidence assigned to the selected response, and the mean confidence conditionalized on whether the response was correct vs. incorrect. The conditionalized confidence and the proportion of correct (vs. incorrect) responses were used to create the following four tests: (1) FA-inducing items: This test included items which, when answered incorrectly, yielded the highest confidence (compared to other items). These items also produced little or no underconfidence when answered correctly.
Thus, FAs should be frequent relative to misses. (2) Miss-inducing items: This test included items which, when answered correctly, yielded the lowest confidence (compared to other items). These items also produced little or no overconfidence when answered incorrectly. Thus, misses should be frequent relative to FAs. (3) Accuracy-inducing items: This test included items which, when answered incorrectly produced low confidence, and when answered correctly produced high confidence. Thus, both FAs and misses should be rare. (4) Half FA-inducing items, half miss-inducing items: This test included an equal number of items from the FA-inducing test and the miss-inducing test described above. Thus, both FAs and misses should be common and should occur with similar frequency.

Each test consisted of 40 items total. Thirty-six of those items were the items that best fit the criteria just described; the remaining four items were neutral (accuracy-inducing) items – the same four items in each test. Because only 80 items were available to create these tests, note that some overlap exists among the items on the tests; however, also note that item statistics fell along a continuum that allowed some items to be appropriate for more than one test. (See Appendix for a list of all items on each test.) Effort was made to keep the expected average performance in each group between 60 and 70% correct. The inclusion of the four neutral items in each group helped to achieve this goal. Expected mean values for percent correct, overall item confidence, and item confidence for correct vs. incorrect responses are shown in Table 4. For comparison to a typical trivia test, the actual mean values for the test used in Study 1 were: 57.1% performance, 6.4 overall confidence, 4.6 confidence for incorrect items, and 7.1 confidence for correct items. Most important, the FA-inducing test was designed to produce higher confidence for incorrect responses (i.e., more FAs) than the other tests; and the miss-inducing test was designed to produce lower confidence for incorrect responses (i.e., more misses) than the other tests.
Table 4. Expected mean values for performance, overall confidence, and confidence conditionalized on response correctness, for each item set (Study 2).

<table>
<thead>
<tr>
<th>Group</th>
<th>Item set</th>
<th>Expected overall performance</th>
<th>Expected overall mean confidence</th>
<th>Expected mean confidence for:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Incorrect responses</td>
</tr>
<tr>
<td>1</td>
<td>FA-inducing</td>
<td>61.9%</td>
<td>7.0</td>
<td>5.1</td>
</tr>
<tr>
<td>2</td>
<td>Miss-inducing</td>
<td>60.1%</td>
<td>5.5</td>
<td>3.7</td>
</tr>
<tr>
<td>3</td>
<td>Accuracy-inducing</td>
<td>68.7%</td>
<td>6.6</td>
<td>3.8</td>
</tr>
<tr>
<td>4</td>
<td>Half-FA/half-miss-inducing</td>
<td>63.7%</td>
<td>6.2</td>
<td>4.1</td>
</tr>
</tbody>
</table>

Notes. Expected values were based on item performance and confidence ratings from previous studies. Expected overall performance is the proportion of times each item was answered correctly in previous studies, averaged across all items in the current item set. Confidence values are on a scale of 0 to 10.

Procedure

The procedure was nearly identical to the procedure used in Study 1. The only difference pertains to measuring item confidence. Specifically, in addition to the item confidence judgments made on a scale of 0-10, participants also explicitly indicated whether they believed their item response was “more likely to be correct” or “more likely to be incorrect.” This additional judgment provided a second method of identifying FAs and misses. All other responses made by participants (i.e., item responses and global estimates) were the same as in Study 1.

Results

Computing frequencies of false alarms and misses

To compute frequencies of FAs and misses, I explored two different methods of identifying high- and low-confidence judgments. The first method (used previously in Study 1) was to apply a midpoint boundary to dichotomize participants’ 0-10 confidence judgments. A second method was to interpret the new, two-alternative confidence judgments (in which participants indicated whether they believed their selected response was “more likely to be
correct” or “more likely to be incorrect”) as indicating high confidence and low confidence, respectively. Unexpectedly, participants exhibited a strong bias towards the more-likely-to-be-correct alternative (vs. more-likely-to-be-incorrect) when judging their two-alternative confidence, thereby exaggerating the appearance of FAs3. In fact, when the frequencies of FAs and misses were based on two-alternative confidence judgments, all four tests produced more FAs than misses (which would suggest that the manipulation had not worked properly); when based on the dichotomized 0-10 judgments, however, this bias was not observed. Thus, the frequencies of FAs and misses in all subsequent analyses were based on dichotomized 0-10 judgments with a midpoint boundary. As in Study 1, I determined the boundary empirically by computing the frequency of high-confidence judgments that would result from each possible boundary and then selecting the boundary that produced the best correspondence between the frequency of high-confidence item judgments and participants’ global estimates (Table 1). The most appropriate boundary was again determined to be the midpoint (i.e., ≥6 denoting high confidence).

Manipulation check

For each of the four tests, I verified that the manipulation had its intended effect on the frequencies of FAs and misses. The FA-inducing test had mean frequencies of 8.3 FAs and 4.4 misses; the miss-inducing test had mean frequencies of 5.1 FAs and 6.2 misses; the accuracy-inducing test had mean frequencies of 5.1 FAs and 5.4 misses; and the half-FA/half-miss-inducing test had mean frequencies of 5.7 FAs and 5.1 misses. Also, the mean levels of performance were similar across the four tests (59.5%, 59.4%, 68.0%, and 65.3%, respectively), but not completely equivalent, $F(3,197) = 5.4, p = .001$. Follow-up comparisons revealed that mean performance on the accuracy-inducing test (68%) exceeded performance on the FA-

---

3 In Study 3, the wording of the alternatives in the two-alternative confidence judgment was altered to reduce bias.
inducing and miss-inducing tests, but no other differences were significant. Most important, performance on the FA-inducing and miss-inducing tests did not differ, indicating that any differences between the estimation accuracy of those groups will reflect differences in global estimates rather than differences in performance. Overall, the manipulation worked approximately as intended.

_Evaluating the item-frequency hypothesis_

First, consider the correspondence between the frequency of high-confidence item judgments and global estimates. The correlation between these two variables was \( r = .69 \) across the four tests (and for each test separately, \( rs = .78, .62, .73, \) and \( .57 \), for the FA-inducing, miss-inducing, accuracy-inducing, and half-FA/half-miss-inducing tests, respectively). Although the correlation was not as strong as in Study 1, it nevertheless was substantial and consistent with the item-frequency hypothesis.

Regarding global estimates, participants’ average global estimates differed by test, \( F(3, 197) = 4.6, p < .01 \). Follow-up comparisons indicated that global estimates on the miss-inducing test (Mean global estimate = 21.1) were lower than the estimates on the other tests (Mean global estimates = 25.6, 26.1, and 25.2, for the FA-inducing, accuracy-inducing, and half-FA/half-miss-inducing tests, respectively). To evaluate the accuracy of these estimates, I computed a difference score (global estimate minus actual global performance) for each individual, such that positive differences represent overestimation, negative differences represent underestimation, and difference close to zero represent accurate estimation. On average, the FA-inducing test produced global overestimation (Mean difference score = 1.78, which was greater than zero, \( t(49) = 2.2, p < .05 \)), the miss-inducing test produced global underestimation (Mean difference score = -2.6, which was less than zero, \( t(49) = -2.5, p < .05 \)), and the remaining two tests produced global estimation accuracy (Mean difference score = -1.1 for the accuracy-
inducing test, which did not differ from zero, \( t(51) = -1.4, p = .18 \); Mean difference score = -.9 for the half-FA/half-miss-inducing test, which also did not differ from zero, \( t(48) = -1.3, p = .20 \). Thus, Study 2 provided clear evidence that the item selection manipulation affected participants’ global estimates and the accuracy of those estimates, which is consistent with the item-frequency hypothesis.

To further test the item-frequency hypothesis, I next investigated the correspondence between the balance of FAs and misses and participants’ global estimation accuracy for performers within each test group. Figure 3 displays the mean frequencies of FAs and misses for each quartile of performers (top boxes in each panel), as well as the global estimation accuracy of each quartile, for each of the four tests. At this level of evaluation, I noted a lack of correspondence between estimation accuracy and the balance of FAs and misses. For example, even though low performers in the FA-inducing test group had, on average, more FAs than misses, \( t(12) = 3.0, p = .01 \), they did not overestimate their performance. However, the most direct test of the hypothesis is the correlation between the imbalance of FAs and misses (i.e., FAs minus misses) and global estimation accuracy. These correlations ranged from .28 to .63 on the four tests, indicating that the imbalance of FAs and misses is a moderately good predictor of global estimation accuracy. Altogether, these results suggest that the item-frequency hypothesis cannot account for all individual differences in global estimation accuracy, but nonetheless it enhances our ability to predict an individual’s estimation accuracy.

Analysis of calibration

Calibration curves were created for each of the four test groups (Figure 4). Looking across these four curves, we can see that the manipulation was reflected in item calibration. Specifically, the FA-inducing test group demonstrated the most overconfidence, whereas the miss-inducing test group demonstrated minimal overconfidence. Also, the accuracy-inducing and
Fig. 3. Mean frequencies of false alarms (FAs) and misses, mean global estimates, and mean global performance, by quartile of performance, for each test group (Study 2).
Fig. 4. Mean calibration of each test group (Study 2).
half-FA/half-miss-inducing test groups demonstrated only small amounts of overconfidence in calibration. Thus, although frequencies are not reflected in calibration curves, the manipulation of confidence could be detected.

**Discrimination accuracy**

For consistency with prior research, item discrimination was computed for each test group and each quartile with the groups. No differences in item discrimination were observed between the test groups, $F(3,197) = 1.8, p = .14$ (*Mean gamma* ranged from .57 to .65). Within the FA-inducing, miss-inducing, and accuracy-inducing test groups, item discrimination differed by quartile of performance (all $p$s < .001), such that lower performers demonstrated poorer discrimination than higher performers (Q1 *gamma* = .47, .65, and .51 for the three tests, respectively; Q4 *gamma* = .73, .78, and .77, respectively). In the fourth test (half-FA/half-miss-inducing), discrimination did not significantly differ by quartile, $F(3,45) = 2.1, p = .12$ (Q1 *gamma* = .50; Q4 *gamma* = .68).

**Discussion**

Study 2 found partial support for the item-frequency hypothesis. In support of the hypothesis, the correlation between the frequency of high-confidence item judgments and global estimates was strong. The average level of global estimation accuracy differed between tests and was predicted by the average frequencies of FAs and misses. Considering quartiles of performers, low performers do appear more susceptible to making FAs, and high performers appear more susceptible to misses, which may be critical for producing the unskilled-and-unaware pattern. However, the imbalance of FAs and misses in a given quartile did not always correspond to the global estimation accuracy of that quartile, suggesting that the hypothesis cannot explain all individual differences in estimation accuracy. Nonetheless, collapsing across
quartiles, the imbalance of FAs and misses was a moderate predictor of global estimation accuracy within the different test groups.

Furthermore, the results indicate that the selection of test items can influence global accuracy and may affect low and high performers differently. As expected, the miss-inducing test produced more underestimation for high performers than low performers. The FA-inducing test was expected to produce the most overestimation for low performers, but this outcome was not observed. One possibility is that the FA-inducing test may not have been sufficiently different from a typical trivia test to produce reliable effects on global estimation accuracy, especially given that the item bank consisted of only 80 different items and some overlap of items occurred between tests. Thus, in Study 3, I aimed to create trivia tests that were more distinct from each other.
STUDY 3

The goal of Study 3 was to evaluate the item-frequency hypothesis with trivia tests that involved a larger distinction between FA-inducing and miss-inducing items. Two tests were created: a FA-inducing test and a miss-inducing test. The items were selected from the same item bank of 80 trivia questions used previously, but each test included only 20 items (rather than 40) that best satisfied the selection criteria, and no items were shared between the tests. The item selection criteria were also modified and will be described in the Method section below.

The tests in Study 3 differed in another way from the tests in Study 2. Namely, the wording of the two-alternative confidence judgment was revised to eliminate the high-confidence bias that was observed in Study 2. Rather than offering alternatives of “more likely to be correct” vs. “more likely to be incorrect,” I asked the participants to select between “probably right” vs. “might be wrong.” This change allowed low-confidence judgments to be selected without requiring participants to be so certain that a selected response was wrong. Assuming that participants faithfully select item responses they perceive as having the best chance of being correct, participants may rarely feel confident that their response is wrong on a multiple-choice test (where correct guessing may easily occur). Thus, the new version of the two-alternative confidence judgment provided another option for identifying FAs and misses.

Method

Participants and Design

Participants were 105 undergraduate students who participated for course credit via the participant pool. The study was conducted in the laboratory. Two different sets of items were constructed to serve as criterion tests (described in the Materials section below), and participants
were randomly assigned to receive one of the two tests: the FA-inducing test (n=53) or the miss-inducing test (n=52).

Materials

Both tests were composed of 20 general knowledge questions drawn from Nelson and Narens’ (1980) norms. The selection of items for each test was based on an item analysis (described below) of the 80 items used in Study 2.

Analysis of items from the previous study. For each trivia item used in Study 2, I computed the proportion of times that it produced a FA and the proportion of times it produced a miss. Based on these proportions, I created the following two tests: (1) FA-inducing items: This test included items which produced FAs a large proportion of the time and rarely produced misses. (2) Miss-inducing items: This test included items which produced misses a large proportion of the time and rarely produced FAs. No accuracy-items group was constructed, since both Studies 1 and 2 already presented examples of tests that yielded balanced FAs and misses and produced global estimation accuracy. Furthermore, because the goal was to create tests that were more differentiated than those in Study 2, the primary focus was on selecting items that produced FAs and misses, rather than on equating performance between the tests.

Procedure

The procedure was nearly identical to Study 2. The only difference pertains to the two-alternative confidence judgments. After each item response and 0-10 confidence judgment, participants were asked to explicitly indicate whether they believed their response was “probably right” or “might be wrong.” Thus, this judgment provided an additional way to identify FAs and misses and was less likely to bias participants towards high confidence (“probably right”) judgments. All other responses made by participants (i.e., item responses and global estimates) were the same as in the previous studies.
Results

*Computing frequencies of false alarms and misses*

Two different methods of identifying high- and low-confidence judgments were explored. In the first method (used in previous studies), I applied a midpoint boundary to dichotomize participants’ 0-10 confidence judgments. The boundary was again determined empirically (Table 1), which resulted in the same boundary being selected (i.e., ≥6 denoting high confidence). In the second method, the two-alternative confidence judgments (in which participants selected “probably right” vs. “might be wrong”) were interpreted as indicating high confidence and low confidence, respectively. These judgments were highly correlated with the dichotomized confidence judgments just described (average gamma correlations were .99 in the FA-inducing test group and .96 in the miss-inducing test group). The frequencies of high confidence judgments computed from these two judgments were also strongly related across participants, \( r = .76 \). In the subsequent analyses, results from both methods will be reported. In general, the two methods yield similar results and lead to the same conclusions.

*Manipulation check*

For both trivia tests, I verified that the manipulation had its intended effect on the frequencies of FAs and misses. The FA-inducing test had mean frequencies of 5.5 FAs and 2.4 misses based on the dichotomized 0-10 judgments (and 4.6 FAs and 2.8 misses based on the two-alternative judgments). The miss-inducing test had mean frequencies of 0.8 FAs and 4.7 misses based on the dichotomized 0-10 judgments (and 0.8 FAs and 5.4 misses based on the two-alternative judgments). Mean levels of performance differed between the two tests (36.6% on the FA-inducing test, 71.3% on the miss-inducing test, \( t(103) = -13.1, p < .001 \)). Overall, the manipulation worked approximately as intended.
Evaluating the item-frequency hypothesis

I first considered the correspondence between the frequency of high-confidence item judgments and global estimates. The correlation between these two variables was $r = .62$ when high confidence was based on the dichotomized 0-10 judgments (and $r = .63$ when based on the two-alternative judgments). Considering the two tests separately, the correlations were $r = .51$ (and $r = .64$) for the FA-inducing test, and $r = .74$ (and $r = .65$) for the miss-inducing test. These correlations indicate correspondence between global estimates and the frequency of high-confidence items judgments, which is consistent with the item-frequency hypothesis.

Participants’ average global estimates did not differ by test, $t(103) = 0.42, p = .67$, but the accuracy of their estimates did, $t(103) = 9.0, p < .001$. Mean difference scores (global estimate minus actual global performance) were +2.45 items (indicating overestimation) for the FA-inducing test group, and -4.8 items (indicating underestimation) for the miss-inducing group. Thus, the item selection manipulation affected the accuracy of participants’ global estimates. The magnitude of the overestimation and underestimation was substantial, especially when considering that the tests contained only 20 items each (10% overestimation and 23% underestimation, respectively).

To further test the item-frequency hypothesis, I investigated the correspondence between the balance of FAs and misses and participants’ global estimation accuracy for performers within each test group. Figure 5 displays the mean frequencies of FAs and misses for each quartile of performers (top boxes in each panel), as well as the global estimation accuracy of each quartile, for both tests. At this level of evaluation, estimation accuracy and the balance of FAs and misses moderately corresponded. Furthermore, correlations between the imbalance of FAs and misses and global estimation accuracy were $.48$ (based on dichotomized 0-10 judgment) and $.62$ (based on two-alternative judgment) in the FA-inducing test group, and $.67$ (dichotomized) and $.59$ (two-
alternative) in the miss-inducing test group. Thus, as in the previous studies, the item-frequency hypothesis was able to account for some of the variance in estimation accuracy.

Fig. 5. Mean frequencies of false alarms (FAs) and misses, mean global estimates, and mean global performance, by quartile of performance, for each test group (Study 3). Note: The values in parentheses are the frequencies based on the two-alternative confidence judgments, whereas the values preceding the parentheses are the frequencies based on the dichotomized 0-10 confidence judgments.

Analysis of calibration

Calibration curves were created for both test groups (Figure 6). As in Study 2, the manipulation was reflected in calibration: Specifically, the FA-inducing test group demonstrated exaggerated overconfidence, whereas the miss-inducing test group demonstrated slight underconfidence.

Discrimination accuracy

For consistency with prior research, item discrimination was computed for both test groups and each quartile within the groups. The discrimination of the miss-inducing test group (Mean gamma = .77) was better than the discrimination of the FA-inducing test group (Mean gamma = .34), t(103) = -8.3, p< .001. Within the test groups, item discrimination did not differ by quartile of performance (ps> .05).
Fig. 6. Mean calibration of each test group (Study 3).

Discussion

Like the previous studies, Study 3 provided some support for the item-frequency hypothesis. Specifically, the correlation between the frequency of high-confidence item judgments and global estimates was strong. The average level of global estimation accuracy differed between tests and was predicted by the average frequencies of FAs and misses. Because these tests differed in performance level, however, differences in global estimation accuracy cannot necessarily be attributed to the manipulation’s effect on estimation levels. Nevertheless, manipulation of item sets is clearly capable of producing the unskilled-and-unaware pattern for trivia questions. Furthermore, low performers appear more susceptible to making FAs, and high performers appear more susceptible to misses. Within test groups, the imbalance of FAs and misses did not perfectly correspond to global estimation accuracy but was a moderate predictor of it. Thus, the item-frequency hypothesis provides some insight into the link between item judgment accuracy and global estimation accuracy and may explain some individual differences in estimation accuracy.
GENERAL DISCUSSION

The present studies confirm that the selection of test items can indeed influence global estimates and their accuracy. Using a bank of 80 trivia questions drawn from Nelson and Naren’s (1980) norms, along with information about participants’ typical levels of performance and confidence for those items, I was able to construct several trivia tests (in Studies 2 and 3) that produced global overestimation or underestimation, in contrast to the estimation accuracy of a typical trivia test (Study 1) which was drawn from the same bank of questions but not manipulated. Importantly, these studies demonstrated that global estimates are indeed linked to item confidence. Further investigating the nature and strength of this link will be important for understanding global estimates and predicting the conditions under which they are inaccurate.

The present studies also evaluated one hypothesis pertaining to the link between item confidence and global estimates. The item-frequency hypothesis proposed that participants construct their global estimates by summing the frequency of high-confidence responses. A corollary of this hypothesis is relevant to the accuracy of global estimates: False alarms (FAs) will contribute to overestimation, misses will contribute to underestimation, and FAs and misses may offset one another. The present studies found moderate support for the item-frequency hypothesis. Specifically, in all three studies, the correspondence between participants’ global estimates and the frequency of high-confidence responses was strong, which is consistent with the idea that the former is based on the latter. Furthermore, when the frequencies of FAs and misses were manipulated in Studies 2 and 3, the average global estimate was affected in the expected direction: An excess of FAs (vs. misses) produced overestimation, and excess of misses (vs. FAs) produced underestimation, and a balance of FAs and misses (whether few or many)
produced accurate estimation. When considering subgroups (quartiles) of test-takers, I also found that low performers generally accumulated more FAs from FA-inducing items than high performers did, and high performers accumulated more misses from miss-inducing items than low performers did, which is consistent with the idea that item selection can produce the unskilled-and-unaware pattern. However, the imbalance of FAs and misses observed within quartiles did not always correspond well with the level of global inaccuracy in that quartile. Thus, the item-frequency hypothesis is limited in its ability to predict individual differences in estimation accuracy and may not entirely capture the basis of participants’ estimates. Below, I speculate about why the item-frequency hypothesis falls short and how it might be modified.

Evidence suggests that people are influenced by their beliefs and motivations when estimating their performance (e.g., Dunning, 1999, Ehrlinger & Dunning, 2003, Norem & Cantor, 1986), but these factors were not incorporated into the item-frequency hypothesis of global estimation. For example, when making a global estimate about test performance, a person is likely to be influenced not only by the confidence she experiences in her item responses, but also by her beliefs about her test-taking ability, about her level of expertise in the test topic, about the how the test will be graded (e.g., partial credit? bonus? on a curve?), and more. These beliefs might be based on past experience with test-taking or other information from the environment. In addition to beliefs, motivations may also influence people’s estimates of performance. For example, a student who feels desperate to achieve a passing grade may give an optimistic estimate, or a student who wants to appear humble or to avoid disappointment when discovering his actual score may give a pessimistic estimate. Beliefs and motivations are important to incorporate in any comprehensive model of global estimation, but they may be difficult to quantify or even identify under what circumstances they are relevant; yet, unless accounted for, they will add error variance to any attempt to link item confidence and global estimates. By
conducting the present studies in a laboratory setting, however, we have likely reduced (though not eliminated) the impact of some of these beliefs and motivations.

Another possible reason for the inadequacy of the item-frequency account is its particular instantiation of the link between item confidence and global estimates. Specifically, the item-frequency account posits that any item response falling above some threshold will be added to a participants’ estimated frequency of correct answers. In this account, an item response with a confidence level of 7 is given the same weight as an item response with confidence level 10 (because they are both above threshold). However, other instantiations of the link between item confidence and global estimation are possible. For example, consider a link that applies different rates of expectation to different levels of confidence. A participant might assume that the majority of their level-10 item responses are correct and should be added to their global estimate, whereas a smaller proportion of their level-7 item responses should be added to their estimate. When investigating other instantiations, researchers might consider the following: How many different levels of confidence do participants actually use and/or monitor? For example, does a single participant typically use all 10 levels provided by the experimenter? If so, does the participant track how many times each separate level was used, or does he track a smaller number of levels (e.g., perhaps a participant would combine some levels of confidence together, thereby reducing the number of confidence levels that need to be monitored). If the participant monitored a small number of combined confidence levels (e.g., high confidence, medium confidence, and low confidence), what percent of responses at each of those levels would the participant expect to actually be correct? In contrast to the expectations defined by researchers using calibration curves, participants might not expect to get 100% of level-10 responses correct or 60% of level-6 responses correct. Consistent with this possibility, Gigerenzer et al. (1991) asked participants to estimate what percent of items will actually be correct at each of 10 confidence levels, and results
indicated that participants did not expect actual rates of correctness to match perfectly with confidence level. Thus, future research may benefit from investigating both (1) how many levels of confidence participants track/monitor, and (2) what rates of correct responses participants actually expect at those various levels. From this perspective, the current research assumed that participants tracked two levels of confidence (high and low) and applied expectation rates of 100% and 0%, respectively. As such, the current studies did not allow for the possibility that participants may anticipate some FAs (e.g., recognizing that high confidence is not a guarantee against mistakes) and anticipate some misses (e.g., expecting some correct guessing, especially on a multiple choice test) and may take these expectations into account when formulating their global estimates. In future studies that do allow for this possibility, FAs may still contribute to overestimates and misses contribute to underestimates if they exceed the levels of FAs and misses that participants expect.

Conclusion

Global estimates and their accuracy are affected by a researcher’s selection of test items. Tests designed to yield many FAs (relative to misses) produced global overestimation, tests designed to yield more misses (relative to FAs) produced underestimation, and tests designed to yield a balance of FAs and misses produced accurate global estimation. Thus, global estimates are linked to item confidence. The item-frequency hypothesis proposed one possible instantiation of this link and was moderately supported in the present studies. Consistent with the hypothesis, the association between participants’ global estimates and the frequency of high-confidence responses was strong. Furthermore, manipulations of item confidence accuracy produced corresponding effects in global estimation accuracy (on average). In these manipulated tests, low (vs. high) performers accumulated more FAs, and high (vs. low) performers accumulated more misses, which may contribute to the unskilled-and-unaware pattern that is sometimes observed.
A main shortcoming of the item-frequency account was that the imbalance of FAs vs. misses did not always predict global estimation within quartiles; but it nonetheless was a moderate predictor across individuals who took a given test. Further research is needed to better clarify the link between item confidence and global estimates.
REFERENCES


Hartwig, M. K. &Dunlosky, J. (under review). Does poor item discrimination produce an unskilled-and-unaware pattern?


## APPENDIX

<table>
<thead>
<tr>
<th>Q#</th>
<th>Trivia Questions</th>
<th>Choices</th>
<th>Study 1</th>
<th>Study 2 Tests</th>
<th>Study 3 Tests</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Typical Test</td>
<td>FA</td>
<td>Miss</td>
</tr>
<tr>
<td>284</td>
<td>WHAT IS THE CAPITOL OF AUSTRALIA?</td>
<td>CANBERRA, SYDNEY, BRISBANE, MELBOURNE</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>WHAT IS THE NAME OF THE SPEARLIKE OBJECT THAT IS THROWN DURING A TRACK MEET?</td>
<td>JAVELIN, DISCUS, SHOTPUT, BATON</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>WHAT IS THE NAME OF THE BIRD THAT CANNOT FLY AND IS THE LARGEST BIRD ON EARTH?</td>
<td>OSTRICH, EGRET, EMU, PENGUIN</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>WHAT IS THE LAST NAME OF THE MAN WHO SHOWED THAT LIGHTNING IS ELECTRICITY?</td>
<td>FRANKLIN, EDISON, JEFFERSON, WATT</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>18</td>
<td>WHAT IS THE NAME OF AN INABILITY TO SLEEP?</td>
<td>INSOMNIA, SLEEP APNEA, NARCOLEPSY, NOCTURIA</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>121</td>
<td>WHAT IS THE CAPITOL OF NEW YORK?</td>
<td>ALBANY, ROCHESTER, NEWYORKCITY, BUFFALO</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>51</td>
<td>WHAT IS THE NAME OF THE POKER HAND IN WHICH ALL OF THE CARDS ARE OF THE SAME SUIT?</td>
<td>FLUSH, STRAIGHT, FULLHOUSE, WHEEL</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>9</td>
<td>WHAT IS THE NAME OF A DRIED GRAPE?</td>
<td>RAISIN, PRUNE, DATE, PLUM</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>47</td>
<td>WHAT IS THE NAME OF A DRIED PLUM?</td>
<td>PRUNE, RAISIN, DATE, AVOCADO</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>5</td>
<td>WHAT WAS THE NAME OF TARZAN'S GIRLFRIEND?</td>
<td>JANE, JAN, JOAN, JUNE</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>74</td>
<td>WHAT IS THE LAST NAME OF THE FIRST PERSON TO COMPLETE A SOLO FLIGHT ACROSS THE ATLANTIC OCEAN?</td>
<td>LINDBERGH, EARHART, WRIGHT, LANGLEY</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>143</td>
<td>WHAT IS THE LAST NAME OF THE ASTRONOMER WHO PUBLISHED IN 1543 HIS THEORY THAT THE EARTH REVOLVES AROUND THE SUN?</td>
<td>COPERNICUS, KEPLER, GALILEO, PTOLEMY</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>53</td>
<td>WHAT IS THE LAST NAME OF THE FIRST PERSON TO SET FOOT ON THE MOON?</td>
<td>ARMSTRONG, GLENN, ALDRIN, SHEPHERD</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>IN WHICH SPORT DOES A RIDER ON HORSEBACK HIT A BALL WITH HIS MALLET?</td>
<td>POLO, LACROSSE, CROQUET, CRICKET</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>66</td>
<td>WHAT IS THE LAST NAME OF THE MYTHICAL GIANT LUMBERJACK?</td>
<td>BUNYAN, GOLIATH, BIGFOOT, RUMPELSTILTSKIN</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Question</td>
<td>Options</td>
<td>Correct Answer</td>
<td>Score</td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------------------------------------------------------------------------</td>
<td>--------------------------------------</td>
<td>----------------</td>
<td>-------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>What is the last name of the union general who defeated the Confederate army at the Civil War battle of Gettysburg?</td>
<td>Meade, Grant, Lee, Custer</td>
<td>X X X X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Who is known as “the father of geometry”?</td>
<td>Euclid, Pythagoras, Archimedes, Descartes</td>
<td>X X X X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>In which type of ski race does the downhill skier make sharp turns around poles?</td>
<td>Slalom, Downhill, Crosscountry, Mogul</td>
<td>X X X X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>What is the last name of Batman’s secret identity in the Batman comics?</td>
<td>Wayne, Parker, Kent, Banner</td>
<td>X X X X X X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>What is the capital of France?</td>
<td>Paris, Versailles, Barcelona, Vienna</td>
<td>X X X X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>In which game are men crowned?</td>
<td>Checkers, Cribbage, Chess, Backgammon</td>
<td>X X X X X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>What is the last name of the man who assassinated Abraham Lincoln?</td>
<td>Booth, Oswald, Ruby, Ray</td>
<td>X X X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>What is the capital of Russia?</td>
<td>Moscow, Minsk, St. Petersburg, Stalingrad</td>
<td>X X X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>What was Frank Lloyd Wright’s profession?</td>
<td>Architect, Writer, Pilot, Doctor</td>
<td>X X X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>What is the last name of the man who proposed the theory of relativity?</td>
<td>Einstein, Planck, Newton, Fermi</td>
<td>X X X X X X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>What kind of a metal is associated with a 50th wedding anniversary?</td>
<td>Gold, Silver, Platinum, Bronze</td>
<td>X X X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>What is the name of Dorothy’s dog in “The Wizard of Oz”?</td>
<td>Toto, Chichi, Bandit, Fido</td>
<td>X X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>What is the name of the thick layer of fat on a whale?</td>
<td>Blubber, Adipose, Pannus, Tallow</td>
<td>X X X X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>What is the name for the astronomical bodies that enter the earth’s atmosphere?</td>
<td>Meteors, Asteroids, Comets, Satellites</td>
<td>X X X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>What is the name of the largest ocean on earth?</td>
<td>Pacific, Atlantic, Indian, Southern</td>
<td>X X X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>What is the name of the organ that produces insulin?</td>
<td>Pancreas, Liver, Kidneys, Gallbladder</td>
<td>X X X X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>What is the last name of the second U.S. president?</td>
<td>Adams, Jefferson, Madison, Monroe</td>
<td>X X X X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>What is the largest planet in our solar system?</td>
<td>Jupiter, Saturn, Uranus, Neptune</td>
<td>X X X X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>WHAT IS THE NAME OF THE OCEAN THAT IS LOCATED BETWEEN AFRICA AND AUSTRALIA?</td>
<td>INDIAN, ATLANTIC, PACIFIC, ARCTIC</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>126</td>
<td>WHAT IS THE NAME OF THE MOUNTAIN RANGE IN WHICH MOUNT EVEREST IS LOCATED?</td>
<td>HIMALAYAS, ROCKIES, APPALACHIANS, ALPS</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>76</td>
<td>WHAT IS THE NAME OF THE LIQUID PORTION OF WHOLE BLOOD?</td>
<td>PLASMA, GLUCOSE, HEMOGLOBIN, WATER</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>11</td>
<td>WHAT IS THE NAME OF THE CRIME IN WHICH A BUILDING OR PROPERTY IS PURPOSELY SET ON FIRE?</td>
<td>ARSON, EXTORTION, FRAUD, RACKETEERING</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>WHAT IS THE NAME OF THE PROCESS BY WHICH PLANTS MAKE THEIR FOOD?</td>
<td>PHOTOSYNTHESIS, OXIDATION, RESPIRATION, FERMENTATION</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>WHAT IS THE NAME OF THE REMAINS OF PLANTS AND ANIMALS THAT ARE FOUND IN STONE?</td>
<td>FOSSILS, PETROGLYPHS, GEOGLYPHS, ARTIFACTS</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>WHAT IS THE LAST NAME OF THE AUTHOR WHO WROTE &quot;ROMEO AND JULIET&quot;?</td>
<td>SHAKESPEARE, CHAUCER, MARLOWE, MACHIAVELLI</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>WHAT IS THE NAME OF THE SEvere HEADACHE THAT RETURNS PERIODICALLY AND OFTEN IS ACCOMPANIED BY NAUSEA?</td>
<td>MIGRAINE, CLUSTER, HEMATOMA THROMBOSIS</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>WHICH PRECIOUS GEM IS RED?</td>
<td>RUBY, SAPPHIRE, AMETHYST, EMERALD</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>WHICH SPORT USES THE TERMS &quot;GUTTER&quot; AND &quot;ALLEY&quot;?</td>
<td>BOWLING, VOLLEYBALL, TENNIS, BILLIARDS</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>10</td>
<td>WHICH SPORT IS ASSOCIATED WITH WIMBLEDON?</td>
<td>TENNIS, GOLF, SOCCER, HORSEBACK</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>63</td>
<td>WHAT IS THE NAME OF THE ISLAND-CITY BELIEVED SINCE ANTIQUITY TO HAVE SUNK INTO THE OCEAN?</td>
<td>ATLANTIS, ELDORADO, BERMUDA, AQUARIUS</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>WHAT IS THE LAST NAME OF THE MAN WHO RODE HORSEBACK IN 1775 TO WARN THAT THE BRITISH WERE COMING?</td>
<td>REVERE, HALE, WASHINGTON, PAINE</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>78</td>
<td>WHAT IS THE NAME OF THE LEGENDARY ONE-EYED GIANT IN GREEK MYTHOLOGY?</td>
<td>CYCLOPS, CERBERUS, CHIMERA, CENTAUR</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>83</td>
<td>WHAT IS THE NAME OF DEER MEAT?</td>
<td>VENISON, VEAL, MUTTON, CHEVON</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>52</td>
<td>WHAT IS THE NAME OF THE NAVIGATION INSTRUMENT USED AT SEA TO PLOT POSITION RELATIVE TO THE MAGNETIC NORTH POLE?</td>
<td>COMPASS, Sextant, ASTROLABE, CHRONOMETER</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Question</td>
<td>Answer Options</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>------------------------------------------------------------------------</td>
<td>----------------------------------</td>
<td>---</td>
<td>---</td>
<td></td>
<td></td>
</tr>
<tr>
<td>What is the last name of the singer who recorded &quot;Heartbreak Hotel&quot; and &quot;All Shook Up&quot;?</td>
<td>Presley, Berry, Haley, Holly</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>What was the last name of the woman who supposedly designed and sewed the first American flag?</td>
<td>Ross, Anthony, Adams, Barton</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>What is the proper name for a badminton bird?</td>
<td>Shuttlecock, Pigeon, Ferry, Cardinal</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>What is the last name of the actress who received the best actress award for the movie &quot;Mary Poppins&quot;?</td>
<td>Andrews, Garland, Charisse, Rogers</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>What is the name of the automobile instrument that measures mileage?</td>
<td>Odometer, Tachometer, Pedometer, Auxometer</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>What is the name of the largest desert on earth?</td>
<td>Sahara, KalaHari, Gobi, Arabian</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>What is the name of the desert people who wander instead of living in one place?</td>
<td>Nomads, Aztecs, Immigrants, Aboriginees</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>What is the only liquid metal at room temperature?</td>
<td>Mercury, Solder, Silver, Antimony</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Who was the leader of the Argonauts?</td>
<td>Jason, Argon, Neil, Caesar</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>What is the term in golf referring to a score of one under par on a particular hole?</td>
<td>Birdie, Eagle, Bogey, Albatross</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>What is the last name of the baseball player who had the most home runs in a single season prior to 1961?</td>
<td>Ruth, Maris, Aaron, Connor</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>What is the name of the city in Italy that is known for its canals?</td>
<td>Venice, Rome, Florence, Milan</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>In what park is &quot;Old Faithful&quot; located?</td>
<td>Yellowstone, Zion, Yosemite, Glacier</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Who was the Egyptian queen who joined forces with Mark Antony of Rome?</td>
<td>Cleopatra, Helen, Cassandra, Octavia</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>What is the word that means a nautical mile per hour?</td>
<td>Knot, Fathom, League, Hullspeed</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>What is the name of the large hairy spider that lives near bananas?</td>
<td>Tarantula, Golliath, Blackwidow, Redback</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>What is the last name of the famous magician and escape artist who died of appendicitis?</td>
<td>Houdini, Knievel, Copperfield, Davenport</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>WHAT IS THE LAST NAME OF THE BOXER WHO LATER BECAME KNOWN AS MOHAMMUD ALI?</td>
<td>CLAY, FRAZIER, FOREMAN, WAGNER</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>------------------------------------------------------------------------</td>
<td>---------------------------------</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>31</td>
<td>WHICH TYPE OF SNAKE DO ASIAN SNAKECHARMERS USE?</td>
<td>COBRA, MOCCASIN, PYTHON, RATTLESNAKE</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>39</td>
<td>OF WHICH COUNTRY IS BUENOS AIRES THE CAPITOL?</td>
<td>ARGENTINA, BRAZIL, CHILE, VENEZUELA</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>10</td>
<td>WHAT IS THE LAST NAME OF THE WOMAN WHO BEGAN THE PROFESSION OF NURSING?</td>
<td>NIGHTINGALE, ADAMS, HENDERSON, BARTON</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>105</td>
<td>WHAT IS THE NAME OF THE SMALLEST JAPANESE STOVE USED FOR OUTDOOR COOKING?</td>
<td>HIBACHI, KILN, WOK, YAKITORI</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>72</td>
<td>IN WHICH EUROPEAN CITY IS THE PARTHENON LOCATED?</td>
<td>ATHENS, ROME, VIENNA, PARIS</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>139</td>
<td>IN WHICH CITY IS HEATHROW AIRPORT LOCATED?</td>
<td>LONDON, CHICAGO, LOS ANGELES, AMSTERDAM</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>153</td>
<td>WHAT IS THE LAST NAME OF THE ARTIST WHO PAINTED &quot;GUERNICA&quot;?</td>
<td>PICASSO, DA VINCI, DALI, VANGOGH</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>233</td>
<td>WHAT IS THE NAME OF THE ISLAND ON WHICH NAPOLEON WAS BORN?</td>
<td>CORSICA, SICILY, CRETE, CYPRUS</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>214</td>
<td>WHAT IS THE LAST NAME OF THE ARTIST WHO PAINTED &quot;GUERNICA&quot;?</td>
<td>PICASSO, DA VINCI, DALI, VANGOGH</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>253</td>
<td>WHAT IS THE LAST NAME OF THE AUTHOR OF &quot;OUR TOWN&quot;?</td>
<td>WILDER, JOYCE, WILLIAMS, MITCHENER</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>160</td>
<td>WHAT IS THE NAME OF THE RIVER ON WHICH BONN IS LOCATED?</td>
<td>RHINE, SEINE, DANUBE, GANGES</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

Note. Question number (in column 1) is used here to label each item; the number is identical to the difficulty ranking assigned in Nelson & Narens’ (1980) norms, such that lower numbers indicated a higher probability of correct recall. The order in which items are listed reflects the selection criteria in Study 2. Specifically, items at the top of the list produced high confidence (on average) for incorrect responses, items at the bottom produced low confidence for correct responses, and items in the middle produced high and low confidence for correct and incorrect responses, respectively. The columns at the right indicate whether the item was included in each test in Studies 1-3. The correct answer to each question is the first choice listed.