What Cognitive Interviewing Reveals about a New Measure of Undergraduate Biology Reasoning

Version accepted 4/25/209 by *Journal of Experimental Education*

Jennifer G. Cromley*[a], Ting Dai[b], Tia Fechter[c], Martin Van Boekel[d],
Frank E. Nelson[e], & Aygul Dane[a]

[a] University of Illinois at Urbana-Champaign ORCID: 0000-0002-6479-9080

[b] University of Illinois at Chicago

[c] Sole proprietor

[d] The SEARCH Institute

[e] Temple University

*Corresponding author

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A160335 to University of Illinois, Urbana-Champaign. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

Abstract

Reasoning skills have been clearly related to achievement in introductory undergraduate biology, a course with a high failure rate that may contribute to dropout from undergraduate STEM majors. Existing measures are focused on the experimental method, such as generating hypotheses, choosing a research method, how to control variables other than those manipulated in an experiment, analyzing data (e.g., naming independent and dependent variables), and drawing conclusions from results. We developed a new measure called Inference-Making and Reasoning in Biology (IMRB) that tests deductive reasoning in biology outside of the context of the experimental method, using not-previously-instructed biology content. We present results from coded cognitive interviews with 86 undergraduate biology students completing the IMRB, using within-subjects comparisons of verbalizations when questions are answered correctly vs. answered incorrectly. Results suggest that the IMRB does in fact tap local and global inferences, but not knowledge acquired before the study or elaborative inferences that require such knowledge. For the most part, reading comprehension/study strategies do not help examinees answer IMRB questions correctly, except for recalling information learned earlier in the measure, summarizing, paraphrasing, skimming, and noting text structure. Likewise, test-taking strategies do not help examinees answer IMRB questions correctly, except for noting that a passage had not mentioned specific information. Similarly, vocabulary did not help examinees answer IMRB questions correctly. With regard to metacognitive monitoring, when questions were answered incorrectly, examinees more often noted a lack of understanding. Thus, we present strong validity evidence for the IMRB, which is available to STEM researchers and measurement experts.

Keywords: biology, reasoning, cognitive interviews

What Cognitive Interviewing Reveals about a New Measure of Undergraduate Biology Reasoning

Introduction

Reasoning is critical for understanding STEM (Science, Technology, Engineering, and Mathematics) content, but few biology-specific reasoning measures exist for use with undergraduate students. While many different skills are required to be successful in introductory biology courses, drawing accurate conclusions from presented material is one important skill, as neither instructors nor authors make all relations explicit (Cromley, Snyder-Hogan, & Luciw-Dubas, 2010b; Otero, Leon, & Graesser, 2002). For example, a typical instructional presentation will state that 1. DNA strands are connected by weak hydrogen bonds and that 2. DNA strands need to pull apart in order for the protein-formation process to begin, but the learner needs to conclude for him- or herself that 3. the weakness of the hydrogen bonds is what allows the strands to complete the necessary step of pulling apart. Numerous studies have shown that when students do not make these inferences, they show poor understanding of the material (Cromley et al., 2010b; Ozuru, Dempsey, & McNamara, 2009; Van Den Broek, Virtue, Everson, Tzeng, & Sung, 2002). Students who do not draw these conclusions may have both pieces of information (1. and 2.) in memory but never connect them by making the inference (3.). Not having made this inference can then interfere with understanding material presented later (e.g., Why are strands of antibody molecules connected with strong disulfide bridges?). Clearly, other types of reasoning are involved in forming a scientific understanding in science learning contexts as well as in scientific practice (e.g., inductive reasoning from observations in nature, hypothesis generation from theory), but the evidence shows that deductive reasoning is clearly involved in understanding science content. There is evidence that prior deductive reasoning scores are related to both students' later academic performance in introductory science courses and their later retention in science majors (Dai & Cromley, 2014). In addition to other predictors of STEM achievement such as high school GPA (e.g., Hernandez et al., 2013), standardized test scores such as SAT or ACT (e.g., Zusho, Pintrich, & Coppola, 2003), and high school class rank (e.g., Martin, Montgomery, & Saphian, 2006), deductive reasoning scores explain additional variance. Furthermore, the same cognitive predictors of undergraduate STEM achievement have also been found to be predictors of undergraduate STEM retention (Maltese & Tai, 2011).

In the present study, we use cognitive interviewing (think-aloud) data, analyzed using a within-subjects design, to provide strong evidence in support of several inferences[1] in an argument to support validity for a new measure of biology reasoning called Inference Making and Reasoning in Biology (IMRB). This validity evidence was gathered from one sample as part of a larger body of data collected from multiple samples, in a program of research based on Kane's (2013) argument-based approach to validity. Specifically, in the present study we report on evidence gathered with the aim of providing supporting evidence for three specific assumptions related to the following inferences (see Appendix for the full range of evidence being gathered):

1. Supporting the description inference[2]: Item development was informed by think-alouds on passages that provided student-formed inaccurate inferences used as response alternatives for multiple-choice items.

[1] For clarification, the term "inference" is used for two purposes in this study. The first is as a measured construct for the IMRB. The second usage stems from using a validation argument approach to supporting the use of the IMRB for its intended purposes. Within a validity argument for an assessment, inferences are generated which have associated assumptions. Those assumptions are then supported through the evaluation of research study results, existing literature, and reasoned semantic arguments.

[2] Description inference: The IMRB consists of clearly defined and developed measurement targets.

2. Supporting the explanation inference[3]: Do cognitive interviews reveal (a) that correct responses are more likely if accurate inference strategies are expressed? and (b) that correct responses are NOT more likely if test-based strategies are used?

3. Supporting the extrapolation inference[4]: Do cognitive interviews reveal that stimuli and items do elicit inference-making skills?

The importance of deductive reasoning in biology. There is a definite need to better understand students' ability to engage in deductive inference-making and reasoning, because as with all learning at the undergraduate level, instructors and textbooks do not make every relation in the domain explicit. In other words, students are required to draw their own conclusions (i.e., engage in inference-making) in order to fully understand the course material. Such inferences are critical for a deep understanding of course material (e.g., Cromley, Snyder-Hogan, & Luciw-Dubas, 2010a). Failure to draw such inferences is directly associated with poorer course grades and lower persistence in STEM majors among students taking these biology courses (e.g., current or future majors in biology, biochemistry, ecology, conservation biology, biotechnology). Furthermore, lower biology grades are related to lower persistence in STEM majors among students required to take these biology courses (Maltese & Tai, 2011), so poor deductive reasoning skills can have detrimental direct and indirect effects on persistence. This highlights the importance of developing high quality measures that can be used to assess deductive reasoning skills.

Another measure of reasoning in biology—Lawson et al. To date, there is a scarcity of published assessments measuring inference-making and reasoning in biology. The most-often

[3] Explanation inference: The expected score on the IMRB can be used to make classification decisions such as identifying examinees that are "at risk" of dropping out of STEM-related fields of study.

[4] Extrapolation inference: The classification of "at-risk" can be interpreted to mean that the examinee lacks an appropriate level of inference-making skill for their STEM-related coursework.

used measure in biology was developed by Lawson and colleagues (e.g., Lawson, Banks, & Logvin, 2007), and uses biology scenarios to test aspects of the scientific method, such as experimental design, graphing, and inferring causality from experiments. A minority of the Lawson et al. items test application of biological principles to specific situations, and these rely on background knowledge and multi-step reasoning from such knowledge. The measure is also rooted in a theoretical framework of cognitive development (Piaget, 1952), with different ranges of scores classified in one of three Piagetian stages: concrete operational, formal operational, and postformal (Lawson et al., 2007, p. 712). Although content on the scientific method is part of gateway biology courses, the emphasis of such courses is *reasoning with biological principles* such as evolution, succession, cell signaling, alternation of generations, and so on. The focus of this study is to present results from one sample collected as part of a research program to establish validity evidence for use of the Inference Making and Reasoning in Biology (IMRB) measure with undergraduate biology students, and highlights two especially critical pieces of evidence supporting explanation and extrapolation inferences gathered from this sample.

Construct definition. The IMRB is *specifically rooted in deductive reasoning about and with biological principles with newly-presented information*, such that the meaning of specific biological terms does not need to be known. That is, even if the student makes a guess at what the biological term means (e.g., lymphocyte), she or he can still reason through the answer to the question. This situation—having a tenuous grasp on new material, but having to reason with it—may in fact characterize much of learning in gateway STEM courses. Evidence to support the explanation inference (i.e., an IMRB score can be used to identify students “at-risk” of dropping out of STEM-related programs of study) for the IMRB must, therefore, show that the measure does in fact tap deductive reasoning and not, for example, vocabulary (i.e., knowledge of word meanings). One way to provide such evidence is to ask examinees to say what they are thinking while they answer IMRB questions, which typically is referred to as cognitive interviewing (Leighton, 2017).

Aims. The data presented in this paper are focused on evidence provided by cognitive interviews to support one assumption associated with the explanation inference and one assumption associated with the extrapolation inference for the IMRB. This paper does not provide the full validity argument for the uses of the IMRB (the framework for the full validity argument can be found in the Appendix, where evidence collected from the present sample is shown in **bold**, and evidence collected from other sources or planned to be collected from other sources is show in grey). Instead, it focuses on two research questions whose answers are intended to provide information to support two specific assumptions associated with their respective inferences for the IMRB.

Research questions. For the study reported here, we posed two specific research questions:

1) What do biology student examinees verbalize (background knowledge, inferences, low- and high-level reading comprehension strategies, metacognitive monitoring, test-taking strategies, vocabulary) while they answer biology reasoning questions on unfamiliar content?

2) What differences are seen in the patterns of verbalizations between correctly-answered and incorrectly-answered IMRB questions?

An examination of answers to these two questions help evaluate three specific validity framework assumptions shown below (A-C). These three assumptions will be referred back to throughout this paper as a way to focus readers' attention on how each analysis and their interpretations are relevant to the larger validity argument for the IMRB.

A. In support of the description inference: Assessment tasks provide evidence of observable attributes in relevant content areas.

B. In support of the explanation inference: Observed test performance is related to the theoretical construct of deductive reasoning with newly-presented information. This implies that observed test performance is not related to other constructs such as test-based strategies.

C. In support of the extrapolation inference: Scores are related to real-world enactment of the assessed skill (i.e., actually drawing the inferences)

The following sections provide some more background on the IMRB assessment creation and existing literature that helped shape our research questions and validity claims.

Creating the IMRB assessment: Supporting the description inference. The IMRB multiple-choice items were created from correct inferences (correct answers) and reasoning errors (distractors) from think-aloud protocols collected from 91 undergraduate biology students who were asked to "read this passage as if you were learning the material for Biology 101" (Cromley et al. 2010a, p. 63). For the IMRB, brief passages were pulled from the longer text used in that study, and questions were written that required inferences within the text passages chosen, either from adjacent sentences or across multiple sentences within the same passage. The questions were designed to be answerable directly via correct deductions within the passage, not to require intermediate strategies such as summarizing, nor to require specialized prior knowledge or specialized vocabulary. It is this design process that provides preliminary support for the description inference—that the stimuli and items are developed to elicit the observable attributes and that the assessment tasks provide evidence of observable attributes in relevant content areas.

An appropriate tool for evaluating assumptions B and C, and thus responding to our research questions is cognitive interviewing, discussed next.

Cognitive interviewing in support of the explanation and extrapolation inferences. Cognitive interviews have been used to provide evidence—especially of explanation inferences—for various types of assessments, ranging from motivational measures (Koskey, Karabenick, Woolley, Bonney, & Dever, 2010) to language tests (e.g., Rupp, Ferne, & Choi, 2006). Cognitive interviews are similar to think-aloud protocols in that participants verbalize their thinking while they complete a task; but cognitive interviewing is used to provide validity evidence for an assessment rather than to provide information about some other type of task such as reading or problem solving (Willis, 2018). Although most cognitive interviewing has been conducted with non-cognitive measures, a number of studies have investigated students completing cognitive assessments. One prominent finding from this literature is that students typically approach answering test questions using test-taking strategies rather than from a perspective of understanding the materials and then answering questions based on that understanding (Rupp et al., 2006). This undermines the explanation inference that observed test performance is related to the theoretical construct (rather than to sources of construct-irrelevant variance). When test-taking strategies such as process of elimination are the primary reason that examinees are able to respond correctly to multiple-choice test questions, the test question (and perhaps the test) has failed at its ability to elicit the desired cognitive ability targeted for measurement (e.g., inference-making). Therefore, it is prudent to uncover the rationale for "why" examinees arrive at a correct response to support both explanation (was the item answered correctly because a correct deductive inference was made?) and extrapolation (did the items indeed elicit deductive inferences?) inferences about an assessment.

Potential threats to the explanation inference. To provide evidence in support of the explanation inference—that observed test performance is not related to constructs other than deductive reasoning—we needed to identify other potential sources of variance (in this case, construct-irrelevant variance) for answering passage-based questions.

Better student performance on the IMRB is expected to result primarily from making correct inferences within the presented material (primarily in support of the explanation inference, but also in support of the evaluation inference[5]). However, given that part of the task examinees performed was reading passages, we anticipated the need to check that other reading-related processes or study strategies were not in fact responsible for observed performance on the IMRB. Based on the literature, potential threats to the explanation inference would be correct answers derived via reading comprehension strategies or study strategies, vocabulary, prior biology knowledge (background knowledge), or test-taking strategies (e.g., Smith, 2017). We also might expect differences in metacognitive monitoring (e.g., 'I think I get it' or 'It's not making sense') between questions answered correctly and questions answered incorrectly. We briefly review these constructs below, as they form the basis for our cognitive interview coding scheme that provides evidence in support of the explanation inference.

Inference/deductive reasoning. If instructors and textbook authors made every relation in the domain explicit, lectures and texts would be substantially longer than they are. Students must therefore actively engage in inference-making in order to form an accurate and well-integrated mental model of course material, above and beyond lists of disconnected facts. Inferences are critical for truly understanding course material rather than simply memorizing information, and play a vital role in transfer of undergraduate biology learning to new contexts (Cromley et al., 2010a, 2010b). Furthermore, science text requires readers to make more inferences, and also inferences that are more difficult, compared to narrative text of the same

[5] Evaluation inference: Observed performance on the IMRB is adequately transformed into a test score that can be used to represent the examinee's ability to make inferences.

length (Van den Broek, 2010). Poor deductive reasoning skills can have detrimental direct effects on persistence in STEM majors, as well as indirect effects via poor grades (i.e., poor reasoning is associated with poor grades and poor grades are associated with dropout; Lawson et al., 2007). In much of the literature, inferences have been divided into logical conclusions made in nearby sections of text (e.g., within a single paragraph), typically called bridging inferences, and inferences made from prior knowledge connected with the current section of text (typically called elaborative inferences; McNamara, 2004). Although bridging inferences are considered to be easier than elaborative inferences, both are considered equally important for forming an accurate and complete mental model of course material.

Background knowledge. The role of background knowledge in reading comprehension has also been widely supported in empirical research and consequently in major theories such as Kintsch's (1998) construction-integration (CI) model. Indeed, the role of background knowledge in reading comprehension has been studied since the "cognitive revolution," and has been researched among middle school (Dole, Valencia, Greer, & Wardrop, 1991), high school (Alexander & Kulikowich, 1991), and undergraduate (McNamara, Levinstein, & Boonthum, 2004) students. With undergraduate science students, Cromley and colleagues (2010b) found that background knowledge and reading vocabulary accounted for the largest variance in text comprehension after accounting for inferences, strategy use, and word reading. In addition to its role directly in text comprehension, background knowledge is related to strategy use such that the reader must have some knowledge of information to use effective strategies (Bonner & Holliday, 2006; Surber & Schroeder, 2007; Taboada & Guthrie, 2006; van den Broek & Kendeou, 2008). Background knowledge is also critical for making accurate elaborative inferences (Gilabert, Martínez, & Vidal-Abarca, 2005; van den Broek & Kendeou, 2008).

Strategies—high-level, low-level, and metacognitive monitoring. There is a large and robust literature on cognitive and metacognitive monitoring strategy use across undergraduate study strategies (e.g., Flippo & Caverly, 2009) and reading comprehension strategies (e.g., Cho & Afflerbach, 2017). Such strategies are consciously applied with the specific goal of improving comprehension and learning. In much of this research, a separation is made between higher-level strategies that transform the information being studied or read about (e.g., Cho & Afflerbach, 2017) and lower-level strategies that do not transform the information. In general, low-level strategies such as highlighting or paraphrasing are not associated with good comprehension or performance, whereas higher-level strategies are associated with better comprehension or performance.

Generally-ineffective low-level strategies include paraphrasing, highlighting, underlining, copying, and re-reading. Generally-effective high-level strategies include summarizing, taking non-verbatim notes, outlining, drawing/sketching, and noticing text structure features such as paragraph headings. Metacognitive monitoring—noticing one's level of comprehension—includes positive statements about comprehension and negative statements about noticing non-comprehension. Both can be associated with better learning, given that failing to notice that one has not understood precludes using fix-up strategies (Ehrlich, Remond, & Tardieu, 1999).

Not only are high-level and metacognitive monitoring strategies associated with better comprehension, but a greater variety or richness of strategies is also associated with better comprehension (Cromley & Wills, 2016). Some models posit that adaptive use of strategies is reflected in adaptive in-the-moment strategy choices, and is a hallmark of high-level comprehension. Thus, we can hypothesize that correctly-answered IMRB questions will be associated with a larger variety of strategies, compared to incorrectly-answered IMRB questions, if we do in fact have evidence to support the explanation inference.

Test-taking strategies. A few recent studies have shown that even when test questions are designed to measure higher-order thinking, test-taking strategies seem to be prevalent reasons for correct responses even among highly-skilled examinees. We summarize three illustrative examples of researchers using think aloud studies to uncover this pattern. Smith (2017) found that for 66% of correct item responses provided by 27 AP History high school students to four National Assessment of Educational Progress (NAEP) Grade 12 History test questions intending to measure Historical Analysis and Interpretation, the actual attributable reason for the correct response was the test-taking strategy—process of elimination—not the skills related to the desired construct. Participants never articulated the historical thinking required to analyze or interpret the information provided on the assessment. Likewise, Rupp, Ferne, and Choi (2006) reported that multiple-choice tests of reading comprehension elicited problem-solving behaviors (i.e., test-taking strategies) such as keyword matching and a combination of process of elimination using prior knowledge and reasoning and guessing among the remaining choices rather than reading comprehension. Again, the reading comprehension processes were not demonstrated by examinees despite being able to find correct answers. Finally, Reich (2009) explored two 10th grade history multiple-choice items that were especially discriminating between high and low performers to uncover response processes for thirteen students and found that the reasons for correct responses were not due to greater facility with the social studies standards addressed on the assessment, but rather due to greater "facility with analytical skills related to narrative, semantic relationships and the printed word" (p. 347) – a literacy skill. Interestingly, test-wiseness—which includes all of these test-taking skills—has been found to correlate positively with verbal ability (Farr et al., 1990). Unfortunately, in all three of these studies, even though test items appeared to be functioning well based on statistical data and standard alignment evidence, think alouds revealed that interpreting scores should be done with caution as construct-irrelevant test-wiseness or literacy skills factors may have been the unintended focus of measurement, and therefore pose a threat to supporting assumptions related to the explanation inference.

One limitation of studies that examine students' test taking processes is that participants are often divided into high-scoring and low-scoring groups whose strategies are then compared (e.g., Heist, Gonzalo, Durning, Torre, & Elnicki, 2014; Hong, Sas, & Sas, 2006). Thus, processes used when questions are answered correctly are mixed with those used when questions are answered incorrectly by the same examinee. By contrast, in the present study we use a within-subjects comparison of processes used on all questions answered correctly versus processes used on all questions answered incorrectly by the same student to provide evidence that the observed test performance is related to the theoretical construct of interest and not related to other constructs.

Vocabulary. The role of vocabulary (i.e., knowledge of word meanings) in reading comprehension has been widely investigated both empirically (e.g., Ash & Baumann, 2017) and theoretically (Perfetti, 2007). According to Ouellette (2006) and Cain and Oakhill (2014), the pivotal role of vocabulary knowledge in reading comprehension not only depends on the breadth of vocabulary (i.e. the number of words that students know), but also on the depth of vocabulary (i.e. what students know about the word, such as multiple meanings of a single word). These two aspects considerably influence students' reading skills, especially higher level comprehension (Cain & Oakhill, 2014). Vocabulary knowledge plays an important role in high school and college students' reading comprehension. Even though single word meanings likely only contribute to sentence-level comprehension, they are a critical prerequisite for enacting high-level strategies and inferences (especially elaborative inferences; Cromley & Azevedo, 2007; Cromley, Snyder-Hogan, & Luciw-Dubas, 2010b). A myriad of other studies support the effect of vocabulary on comprehension as mediated by strategies and inference (e.g., Ahmed et al., 2016; Daugaard, Cain, & Elbro, 2017; Segers & Verhoeven, 2016).

Expectations for cognitive interviewing in support of the explanation inference. Based on prior research reporting on cognitive interviewing on test questions, we expect that examinees will use test-taking strategies such as ‘process of elimination,’ but the extent of use on the IMRB will not differ between questions answered correctly and questions answered incorrectly. If this pattern holds for the IMRB, it will suggest that test-taking strategies do not enhance students’ ability to answer our deductive reasoning questions correctly, even though they are used frequently by students when answering test questions. Put another way, it will support our assumption that the use of test-based strategies alone does not result in increased likelihood of correct responses. A positive influence of test-taking strategies on IMRB scores would be classified as construct-irrelevant variance and thereby a threat to the explanation inference. It should be clarified here that some strategies, such as process of elimination, can be used in construct-relevant ways. For example, if an examinee eliminated multiple-choice options based on informed deductive reasoning, that use of the strategy was coded as making an inference and not as using process of elimination. However, if, for example, the examinee eliminated multiple-choice options based on how long the options were, that would reflect a form of test-wiseness and would not be aligned to the intended construct; and thus was coded as using process of elimination.

For our assessment of reasoning with newly-presented information, we expect that when students answer an IMRB question correctly there will be more evidence that they used correct deductive reasoning than when students answer a question incorrectly. This finding would provide evidence in support of the explanation and evaluation inferences, that correct responses result from the use of accurate inference-based strategies. In turn, support for this assumption would suggest that the assessment does indeed tap reasoning with new information, which is the construct we intend to assess.

If our measure in fact taps inference-making skills, and not background knowledge or use of test-taking strategies, we would expect correct answers to be associated with use of global or local inferences, and not with verbalizations of prior knowledge, inferences that incorporate prior knowledge (i.e., elaborative inferences) or vocabulary, all in support of the explanation inference. If we find that inferences are actually verbalized while completing the IMRB, we will have evidence in support of the extrapolation inference.

Finally, with regard to reading comprehension strategy use or study strategy use, we expect no difference in strategy use when answering correctly versus incorrectly, but there is a possibility that examinees will accurately enact some cognitive strategies (e.g., summarizing) as precursors to making accurate inferences (Cromley & Wills, 2016), which might yield more high-level strategy use on questions answered correctly, despite the measure not requiring strategy use.

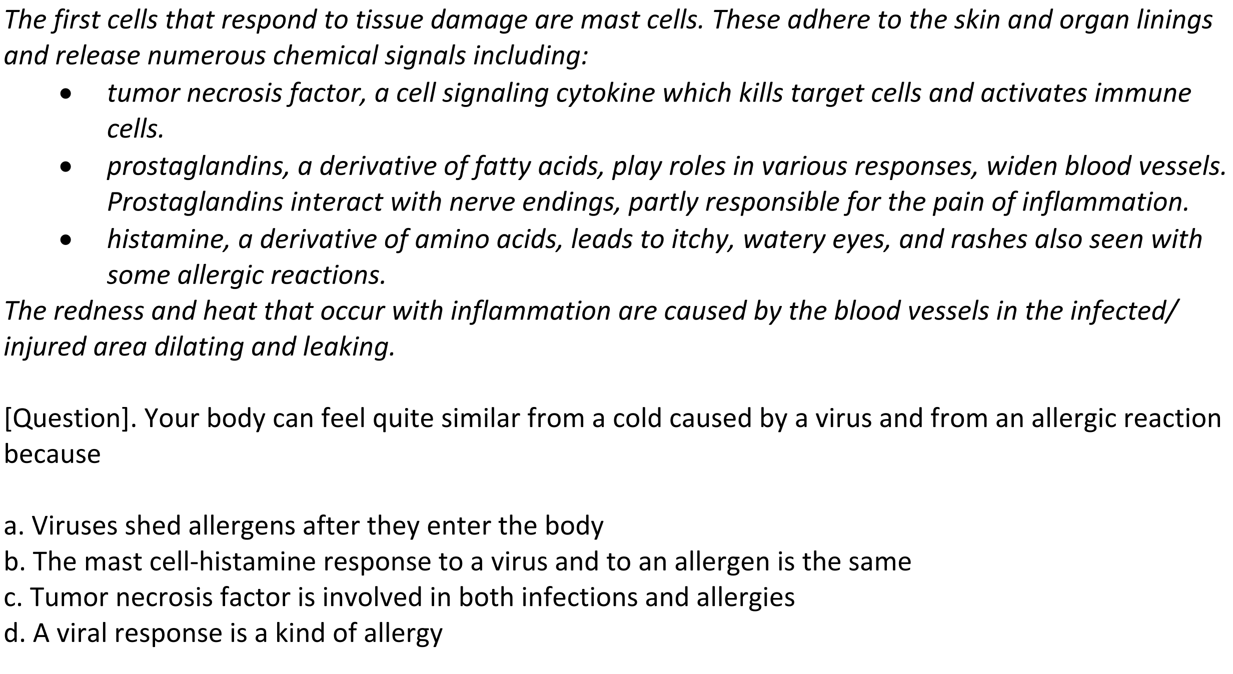
Method

Participants. Eighty-six examinees were recruited with an open invitation to students who had completed an introductory environmental and organismal biology course required for life sciences majors within the past 2 years (*M* = 2.6 semesters prior to the cognitive interview, *SD* = 1.2). Examinees were recruited from two US universities and were paid $35 each for their 1-hour participation. Their mean age was 20.1 (*SD* = 1.2), and they were mostly Sophomores (33%) and Juniors (59%). They were 51% White, 37% Asian, 7% Latino/Latina, and 6% of other races. Twenty-one percent of examinees were first-generation (neither parent with a Bachelor's degree) college students.

Materials. Examinees answered 15 multiple-choice questions that asked them to reason about information related to the immune system. Short paragraphs presented the new information, and all questions were in a 4-option format, answered on a computer using a study-specific Blackboard site (see Figure 1 for an item analogous to those used in the study). Previous unpublished analyses have shown the measure to be unidimensional, to show reliability > .7, to significantly correlate with course grades, to have a range of item difficulties, and to show no bias (as assessed by Differential Item Functioning) by race, sex, or first-generation college status. The measure explains a mean additional 13.5% variance in final introductory biology course grade after accounting for SAT reading and math scores.

Figure 1

Example of a Biology Reasoning Item



The first cells that respond to tissue damage are mast cells. These adhere to the skin and organ linings and release numerous chemical signals including:

- *tumor necrosis factor, a cell signaling cytokine which kills target cells and activates immune cells.*
- *prostaglandins, a derivative of fatty acids, play roles in various responses, widen blood vessels. Prostaglandins interact with nerve endings, partly responsible for the pain of inflammation.*
- *histamine, a derivative of amino acids, leads to itchy, watery eyes, and rashes also seen with some allergic reactions.*

The redness and heat that occur with inflammation are caused by the blood vessels in the infected/ injured area dilating and leaking.

[Question]. Your body can feel quite similar from a cold caused by a virus and from an allergic reaction because

a. Viruses shed allergens after they enter the body
b. The mast cell-histamine response to a virus and to an allergen is the same
c. Tumor necrosis factor is involved in both infections and allergies
d. A viral response is a kind of allergy

Procedure. In an individual 1-hour session in fall 2016-spring 2017, examinees gave written consent and permission to be audio-recorded, audio recording was begun, they were asked to say everything they were thinking as they answered the questions (about 45 min), and they then completed a demographic form. The rationale for using a think-aloud approach to cognitive interviewing was to interfere as little as possible in the examinee's thinking process (Willson & Miller, 2014). Moreover, data were collected as the examinee took the IMRB—rather than retrospectively—so as to capture reasoning in real time, which is the focal construct. In addition, retrospective reports represent retrieval from long-term memory—memory of the answered question—rather than working memory, i.e., the process of problem solving (Leighton 2017). All examinees completed the 15-item measure within the time allotted. Examinees had paper, a pencil, and a pen available if they chose to use them; no other resources were permitted (these are the same conditions under which the IMRB is administered).

Our cognitive interview directions were as follows: "You are being presented with multiple-choice questions based on material taught in [introductory biology]. We are interested in learning about how examinees think about and answer questions such as these, in order to improve the measure we have created. I want you to answer the questions as if you were being tested on them in a course such as [course name]. There is no time limit for answering the questions, but we expect it may take you about 25 minutes to answer them. In this experiment I am interested in what you are thinking as you answer questions. In order to do this I will ask you to THINK ALOUD as you are reading and answering the questions. What I mean by think aloud is that I want you to say out loud EVERYTHING that you are thinking from the time you start reading each question until you give an answer." If the examinee was not verbalizing, we prompted with one of two prompts "Say what you are thinking" or "Say what you are doing" (e.g., if the examinee was taking notes but did not state that). If examinees asked questions, the researcher stated that he or she was not able to answer them.

Two practice questions—a statistics question and one on the history of science—were presented to give examinees a chance to practice thinking aloud (not analyzed). No strategies were suggested or modeled, nor was any feedback on learning processes or correctness of answers given at any point.

Scoring, coding, and data analyses. We first scored examinees' test item answers as correct or incorrect. We then transcribed the protocols verbatim and coded them using a coding scheme based on Cromley et al. (2010a), with 4 added codes related to test-taking strategies (See Table 1 for all codes, specific definitions, and examples from our corpus). These codes correspond to the evidence in support of the assumptions related to the explanation and extrapolation inferences regarding examinee inferences and other strategies as described above. Each codable utterance was assigned one and only one code; uncodable utterances included "OK," "Hmmm," and other content-free statements. The coding scheme includes 7 major categories of cognitive activities, with a total of 41 possible codes:

1) verbalization of specific prior knowledge, either knowledge not presented in the test or knowledge from a prior test passage;

2) inferences, including bridging and elaborative inferences and hypotheses/predictions;

3) low-level reading strategies such as skimming and paraphrasing;

4) high-level reading comprehension strategies such as summarizing, self-questioning, taking notes;

5) metacognitive monitoring strategies;

6) test-taking strategies such as process of elimination; and

7) verbalizations about understanding the meaning of vocabulary words.

Table 1

Coding scheme

Type	Code	Definition	Example
Prior knowledge activation	Prior Knowledge Activation from Before Study PKABS + for accurate - For inaccurate	Participant mentions some specific background knowledge fact (not as part of an inference)	"I'm assuming it's talking about MHC II and I proteins"
Prior knowledge activation	Prior Knowledge Activation from Earlier in Text PKAET + for accurate - For inaccurate	Participant mentions some specific background knowledge fact (not as part of an inference)	"There was something about it has to get activated"
Inference	Hypothesis HYP	Pose a hypothesis about how something might work (the hypothesis is not stated in the text)	"Like fungi could all be the same one…the cells could all be used on the whole group"
Inference	Inference Global INFGLOB + for accurate - For inaccurate	Participant makes a generalization across a large segment of text (not just a summary)	So helper T cells help out all the other WBCs: cytotoxic T cells, B cells, memory cells
Inference	Inference Local INFLOC + for accurate - For inaccurate	Participant makes a conclusion across 2 adjacent sentences	"The secondary immune response that's provoked by the primary response"
Inference	Knowledge Elaboration Before Study KEBS + for accurate - For inaccurate	Participant adds information not in text + info from text and draws a conclusion	"The lymphocytes are good, so I don't think we would want to stop cloning" "They have like the same….conformational compositions…so that would

Type	Code	Definition	Example
			mean the ligand binding site would probably work for that”
Inference	Knowledge Elaboration Earlier in Text KEET + for accurate - For inaccurate	Participant adds information read in a previous passage + info from current text and draws a conclusion	“it didn’t activate anything but future infections is why the memory cells are there for sure”
Low-Level Strategy	Importance of Information IOI	Statement that some information is important	This part is important
Low-Level Strategy	Paraphrase PARA + for accurate - For inaccurate	Participant re-states information from within 1 sentence (not re-reading)	“they are specific for normally just like one foreign type of pathogen which they can interact with”
Low-Level Strategy	Re-read RR	Participant re-reads a segment of text 5 words or longer. Each continuous “chunk” of text without switching to a new activity gets one code, no matter how short or long	“*The ‘selection’ of a lymphocyte by one of the microbe’s antigens… The ‘selection’ of a lymphocyte by one of the microbe’s antigens”*
Low-Level Strategy	Skim SKIM	Reading partial bits of text, in order, rapidly (does not have to say ‘I’m skimming”)	“*consists of memory cells….receptors specific…same antigen’*
Low-Level Strategy	Task Difficulty TD	Statement about how difficult or easy the task is	“That shouldn’t be that bad”
High-level Strategy	Adequacy of Text AT	Statement that text has what is needed	“I know it talks about B, T cells”
High-level Strategy	Coordinating Informational Sources COIS	Compares or matches information between text and diagram (or vice versa)	“Parent lymphocyte becomes effector cells, subgroups become effector cells”

Type	Code	Definition	Example
High-level Strategy	Drawing DRAW	Learner makes his/her own drawing/sketch in notes	"Drawing this out"
High-level Strategy	Evaluating EVAL	Judgment of the learning material (not adequacy or inadequacy of writing)	"So this is kinda dumbed down"
High-level Strategy	Help-Seeking Behavior HSB	Statements about looking to other sources or people for information	"So I would probably ask you know a TA"
High-level Strategy	Inadequacy of Diagram ID	Statement that diagram is confusing, wrong, or inadequate in some other way (not that learner is confused)	This diagram is so bad
High-level Strategy	Inadequacy of Text IT	Statement that text is confusing, wrong, or inadequate in some other way (not that learner is confused)	"OK, this is kinda wordy"
High-level Strategy	Planning PLAN	Participant plans out an approach or a set of steps or stages	"So I'll read the passage first"
High-level Strategy	Search SEARCH	Participant looks in text to find a specific piece of information	"just to see where again it specifically talks about the lymphocytes" "Do they mention macrophages? [then skims]" "So let's look up primary immune response in the passage"
High-level Strategy	Self-Questioning SQ	Participant generates his/her own question	"What is the difference?"
High-level Strategy	Summarize SUM	Participant re-states information from 2+ adjacent sentences	"So it says it recognizes by dividing and memory" "

Type	Code	Definition	Example
	+ for accurate - For inaccurate		
High-level Strategy	Taking Notes TN	Participant states they are taking or took notes (not drawing)	"What I wrote is 2 Abs on each Ig"
High-level Strategy	Text Structure TS	Uses information about how texts are usually structured—bolding, passage-based test questions, order in which info is 'usually' presented	"OK question 8 is also a passage based one"
Metacognitive monitoring	Feeling of Knowing FOK	Verbalizes the feeling that something is understood (not simply "OK" or "Right")	"OK, I remember that much at least"
Metacognitive monitoring	Judgment of Learning JOL	Verbalizes the feeling that something is not understood	"So I feel like that's a bit confusing'
Test-taking	Passage Doesn't Mention It PDMI	Participant notes the lack of mention of specific content	"Cause the passage didn't talk about genes"
Test-taking	Re-Reads Question RRQ[6]	Re-reading any part of the question stem and/or answer options. The 5-words rule does not apply.	"*What is due to genes.* Self means. *What is due to genes*"
Test-taking	Response Strategy-Eliminate RSE	Eliminate an answer option	"And it wouldn't be [D]" "*Cloning*....No"
Test-taking	Response Strategy-Other RSO	Test-taking strategies other than Guessing or Elimination, or evidence of test-taking schemata—includes checking all answer options, following test-taking advice such as choosing a more specific answer over a	"Like I said earlier, I'm going to check over that" "because it's more specific than d"

[6] Re-reading the question could have been placed with low-level strategies, as we do not believe it would produce construct-irrelevant variance. However, it was one of the new test-related codes added to an existing coding scheme that has been previously used with connected text only. We therefore place it in the 'test-taking strategies' category.

Type	Code	Definition	Example
		more general answer, reading the question stem and answer options before reading the passage, noticing that a question format is similar to one presented earlier.	“let’s see the answer choices first [before reading the passage]” “I’ve narrowed it down to b or d” “this answer ***could*** be right” [Neither a JOL nor an FOK, not an RSE] “So it’s the same instructions from earlier” “I’m not even going to read the other responses”
Vocabulary	Vocabulary VOC + for accurate	+ is an accurate paraphrase of the meaning of a word, reading (i.e., …) as ‘which means’	“*memory cells,* which are *long lived cells”*
	- For inaccurate	- is a statement that the meaning is not known, or a search for the definition or an inaccurate paraphrase of the meaning of a word	“lymphocytes was mentioned where?”

Coding yielded a total of 9,705 coded utterances, which were entered into a database with the examinee identifier, question identifier, and flags for correctly-answered and incorrectly-answered questions. Think-aloud studies always face a challenge due to some respondents being more talkative than others, though verbosity is not related to performance. Therefore, as is typical for think-aloud studies to correct for verbosity, we converted each examinee's raw frequency of verbalizations to proportions for each coding category within correctly-answered and incorrectly-answered questions (e.g., an examinee who verbalized accurate global inferences [INFGLOB+] 4 times out of 80 verbalizations in correctly-answered questions used INFGLOB+ at a proportion of .05; an examinee who verbalized failing to understand [judgment of learning;

JOL] 6 times out of 48 verbalizations in incorrectly-answered questions used JOL at a proportion of .13). As is typical for this sort of data, the proportion variables were severely positively skewed for all codes (note the large standard deviations relative to the means in Table 2). This occurs because many participants verbalized a code at a very low rate (e.g., 1-2%), fewer verbalized it at a higher rate (e.g., 3-4%), even fewer verbalized it at higher rates (> 5%). Due to this (expected) non-normality, we proceeded with non-parametric analyses.

For descriptive statistics, we used Spearman rank correlations, and we used a within-subjects analysis (Wilcoxon signed rank test) to compare for each code the proportions verbalized when questions were answered incorrectly to proportions verbalized when questions were answered correctly in order to determine which codes were associated with correct answers. For example, if the mean proportion of inferences on correctly answered questions had on average a significantly higher rank than the mean proportion of inferences on incorrectly answered questions, that would be evidence in support of the explanation inference. We used Cohen's *d* as the effect size metric. All analyses were conducted using SPSS ver. 24.

Results

Descriptive statistics. The mean number of coded verbalizations per examinee was 112.85 (*SD* = 55.65); on average, each examinee made 18.33 (*SD* = 5.06) different types of coded verbalizations.

Research Question 1) What do biology student examinees verbalize (background knowledge, inferences, low- and high-level reading comprehension strategies, metacognitive monitoring, test-taking strategies, vocabulary) while they answer biology reasoning questions on unfamiliar content?

As shown in Table 2, the most commonly coded verbalizations were re-reading questions (RRQ), process-of-elimination (RSE), feeling of knowing (FOK), and re-reading (RR), all of which are rather shallow approaches to the passages and questions. Activating background knowledge comprised a mean 6.8% of verbalizations, inferences comprised a mean of 17.8%, comprehension strategies a mean and 42.8%, test-taking strategies a mean of 45.7%, monitoring a mean of 17.8%, and vocabulary a mean of 7.6%. The presence of inferences in the cognitive interviews supports the extrapolation inference—the passages and items do indeed prompt the use of deductive inferences, which are the assessed skill.

Table 2

Number of Examinees, Count, and Proportion For all Codes, across all Questions

Category/Code	*N*	*Count*	*M* *proportion*	*SD*
Background knowledge				
PKABS-	21	30	1.20%	0.92%
PKABS+	66	191	2.52%	1.74%
PKAET-	9	11	1.03%	0.41%
PKAET+	50	133	2.07%	1.39%
Total		365	6.82%	
Inference				
HYP	6	9	1.18%	0.75%
INFGLOB-	38	64	1.56%	1.40%
INFGLOB+	69	321	3.98%	2.87%
INFLOC-	44	83	1.71%	1.30%
INFLOC+	68	253	3.12%	1.92%
KEBS-	21	34	1.31%	0.94%
KEBS+	52	125	2.03%	1.41%
KEET-	12	14	0.83%	0.50%
KEET+	36	72	1.52%	0.88%
Total		975	17.24%	
Low-level strategies				
IOI	15	24	1.23%	0.96%
PARA-	31	49	1.42%	1.02%
PARA+	83	567	5.83%	3.06%
RR	85	1007	11.43%	8.10%
SKIM	26	55	2.28%	2.10%
TD	23	42	2.25%	3.05%
Total		1744	24.44%	
High-level strategies				
AT	5	5	0.75%	0.43%
COIS	9	11	1.56%	0.95%

			M	
Category/Code	*N*	*Count*	*proportion*	*SD*
DRAW	2	2	0.68%	0.07%
EVAL+	1	1	0.50%	N/A
HSB	1	1	1.01%	N/A
ID	1	1	0.91%	N/A
IT	10	12	1.14%	0.91%
PLAN	25	39	1.41%	1.11%
SEARCH	32	62	1.49%	0.95%
SQ	27	49	1.62%	1.23%
SUM-	17	25	0.94%	0.45%
SUM+	67	195	2.56%	1.85%
TN	3	14	2.51%	2.14%
TS	31	45	1.32%	0.82%
Total		462	18.40%	
Metacognitive monitoring				
FOK	85	1353	13.91%	5.54%
JOL	81	369	3.91%	2.54%
Total		1722	17.82%	
Test-taking strategies				
RRQ	84	1535	16.83%	10.45%
RSE	84	1794	19.60%	8.49%
RSO	73	423	4.69%	3.25%
PDMI	77	402	4.58%	2.53%
Total		4154	45.70%	
Vocabulary				
VOC-	33	92	5.17%	17.12%
VOC+	73	191	2.39%	1.53%
Total		283	7.56%	

Examinees verbalized a range of test-taking strategies coded under the Response Strategy-Other (RSO) code:

1. Not choosing a less-specific option when there was a more-specific option, e.g., “I think it’s B only because it’s more specific than D.” [Examinee 010]
2. Avoiding options with “all,” “never,” “always,” and other such absolutes, e.g., “I don’t think that you can say that *macrophages are white blood cells* because that seems like an absolute um that’s too exclusive.” [Examinee 040]

3. Trying to spot "trick" questions, e.g., "I think A is just kinda thrown in there sometimes to be ...the answer that's kinda a trick." [Examinee 048]
4. Checking all options even after expressing confidence that one option is correct, e.g., "*Therefore, autoimmune diseases are immune responses to self components,* yes, let's read the other options though." [Examinee 043]
5. Reading the question before reading the passage, e.g., "I need to start reading the question before I read these long things." [Examinee 048]

While these test-taking strategies may be generically useful, they sometimes led students to doubt an option that was actually correct, to spend a great deal of time on a relatively straightforward question, or to continue working even after giving good reasoning for a correct answer. For example, examinee 039 stated "So I would think that [the answer] would be D... because that would allow lymphocytes to recognize different types of pathogens" but then continued "And it seems that A, [reading] *pathogens have evolved defenses against the immune system* is kind of irrelevant. Same with C, non-specific and specific defenses, it seems like this one is just talking about specific defenses. So B and D [are still options]. B [reading] *some types of lymphocytes attack all pathogens indiscriminately*, that would appear to be false to me." In this case, the examinee found, was confident in, and stayed with the correct answer but used up time checking the other options.

To give a second example of Response Strategy-Other (RSO), examinee 034 stated "And then D says *your own body; anything that is not part of your own body* so I feel like that's a good option. Um, [C] *what you were born with; what you were not born with,* I don't want to completely rule that one out. [B] *What is due to genes; what is due to environment*, I don't know if they really talk about that but I feel like maybe you could say that, but C and D are stronger choices so I'm gonna rule out B. Hmmm, I might want to go with C, *what you were born with;*

what you were not born with because it talked about how molecules were *already present in the body."* In this case, reading through the options led the examinee away from the correct answer (D) to an incorrect answer (C, which is a known incorrect inference made by biology students).

To give a third example of Response Strategy-Other (RSO), examinee 014 reasoned "*Macrophages phagocytize pathogens. And protein markers on pathogens are called antigens. Therefore, Antigens make antibodies. Antigens phagocytize macrophages.* Um that is not true because macrophages eats pathogens and then antigens are a part of the pathogens, or I mean it's the same thing as pathogens basically so that B is not true. *Macrophages phagocytize antigens.* Um *protein markers* um. Phagocyte means eating a cell not necessarily a protein. Um which is an antigen so protein marker. *D* would be *proteins phagocytize antigens.* Um proteins don't do that thing cuz macrophages are cells. Proteins don't do that um. So most likely it would be wait. *Antigens make antibodies*. And then *protein markers on pathogens are called antigens.* Antigens don't make antibodies. Um although they probably they influence creation of them or like proliferation of them. Um phagocytize ok um. *Antigens make antibodies.* I don't know. So B and C, no B and D are not true, definitely. And then *antigens make antibodies* um. *Protein markers on pathogens.* I mean antigens are like protein markers um they get recognized as nonself cells um. So they do make antibodies cuz antibodies are already proteins. Um and so I guess *macrophages phagocytize antigens* and that it's a part of the pathogen, which is eaten by macrophages." In this case, the examinee made the correct, straightforward inference almost immediately ("macrophages eats pathogens and then antigens are a part of the pathogens"), but took an additional 15 turns before returning to that same inference as the chosen, correct, option.

These examples of Response Strategy-Other (RSO) illustrate how a strategy that might be thought generically useful for taking multiple choice tests could create a disadvantage on our measure of biology reasoning. In sum, even though examinees might have used test-taking strategies (e.g., Response Strategy-Elimination; RSE or Response Strategy-Other; RSO) because they believed these would result in a greater likelihood of responding correctly to test questions, our results suggest that this was not the case. In the case of re-reading the question (RRQ), use of the strategy resulted in a greater likelihood that the student responded incorrectly to the test question. With a small effect size, greater use of the Passage Didn't Mention It (PDMI) code was associated with correct answers compared to incorrect answers (4.4% vs. 3.6%), but it frequently led examinees astray (e.g., when they looked in the passage only for the exact same word as the one in a question). These findings are evidence supporting the assumption that the use of test-based strategies alone does not result in increased likelihood of correct responses; thus providing support for the explanation inference.

Research Question 2: What differences are seen in the patterns of inferences and strategy use when examinees answer questions correctly versus incorrectly?

In the descriptive analyses separately by questions answered correctly versus questions answered incorrectly (See Table 3), it can be seen that the overall patterns are similar, but inferences are more common when questions are answered correctly (17.3% vs. 14.7%), as are high-level comprehension strategies (25.1% vs. 20.3%). Specifically, accurate inferences are more common in correctly-answered than incorrectly-answered questions (13.3% vs. 6.8%).

Results from the within-subjects comparisons for each code are shown in Table 3. First, we predicted that, if the IMRB does in fact measure deductive inferences within passages, there should be significantly more correct inferences in questions answered correctly versus incorrectly. This hypothesis was supported for accurate local inferences (INFLOC+, $p < .01$, Cohen's $d = 0.53$) and for accurate global inferences (INFGLOB+, $p < .01$, Cohen's $d = 1.13$).

Table 3

Within-subjects Count and Percentage of Non-Zero Verbalizations by Correct and Incorrect Answers with Comparisons for Each Code

Code	*z*	*p*	*n* of students verbalizing the code	*n* of times verbalized: answer correct	*M* % of verbalizations within correct answers	*SD*	*N* of times verbalized: answer incorrect	*M* % of verbalizations within incorrect answers	*SD*	*d*
Background knowledge										
PKABS-	.017	.97	21	20	1.1	1.2	10	2.2	5.5	-0.32
PKABS+	1.147	.52	66	135	2.4	2.2	56	2.1	3.1	0.12
PKAET-	.178	.86	9	8	1.0	1.3	3	.8	1.2	0.15
PKAET+	**4.364**	**< .01**	**50**	**112**	**2.3**	**1.6**	**21**	**.9**	**1.7**	**0.86**
PKABS-	.017	.97	21	20	1.1	1.2	10	2.2	5.5	-0.32
Inference										
HYP	1.153	.25	6	8	1.2	1.0	1	.3	.8	0.92
INFGLOB-	1.530	.13	38	37	1.1	1.1	27	2.1	2.8	-0.49
INFGLOB+	**6.242**	**< .01**	**69**	**283**	**4.9**	**3.6**	**38**	**1**	**2.1**	**1.13**
INFLOC-	1.109	.27	44	45	1.1	1.0	38	2.1	2.8	-0.49
INFLOC+	**3.587**	**< .01**	**68**	**201**	**3.5**	**2.3**	**52**	**2.1**	**3.6**	**0.53**
KEBS-	.122	.9	21	16	1.0	1.1	18	1.9	3.4	-0.34
KEBS+	.005	1.00	52	88	2.0	1.9	37	2.7	4.7	-0.23
KEET-	.471	.64	12	8	.8	1.1	6	1.8	2.8	-0.47
KEET+	**3.040**	**< .01**	**36**	**62**	**1.7**	**1.3**	**10**	**.7**	**2.2**	**0.68**
Low-level cognitive strategies										
IOI	1.108	.27	15	16	.9	.9	8	1.7	2.1	-0.55
PARA-	.049	.96	31	28	1.1	1.2	21	1.7	3.1	-0.27
PARA+	**2.960**	**< .01**	**83**	**443**	**6.6**	**5.2**	**124**	**4.7**	**6.0**	**0.35**
RR	.288	.77	85	661	10.6	8.1	346	11.2	9.8	-0.07
SKIM	**2.274**	**.02**	**26**	**41**	**2.3**	**2.2**	**14**	**2.1**	**4.8**	**0.06**
TD	1.232	.22	23	29	2.5	4.2	13	2.0	3.6	0.12
High-level cognitive strategies										
AT	**2.032**	**.04**	**5**	**5**	**1.1**	**.4**	**0**	**0.0**	**0.0**	**N/A**
COIS	.415	.69	9	7	1.8	2.1	4	1.9	3.6	-0.03

Code	*z*	*p*	*n* of students verbalizing the code	*n* of times verbalized: answer correct	*M* % of verbalizations within correct answers	*SD*	*N* of times verbalized: answer incorrect	*M* % of verbalizations within incorrect answers	*SD*	*d*
DRAW	1.342	.18	2	2	.9	.1	0	0.0	0.0	N/A
EVAL	NA	NA	1	1	.7	NA	0	0.0	0.0	N/A
HSB	NA	NA	1	0	0	NA	1	1.9	0.0	N/A
ID	NA	NA	1	0	0	NA	1	2.3	0.0	N/A
IT	1.122	.26	10	10	1.1	1.1	2	.4	1.1	0.63
PLAN	.901	.37	25	28	1.4	1.6	11	1.2	2.0	0.11
SEARCH	.019	.99	32	43	1.3	1.2	19	1.3	1.9	< 0.01
SQ	.048	.96	27	28	1.1	2.2	21	1.4	2.1	-0.14
SUM-	.402	.69	17	15	.9	1.3	10	1.2	1.6	-0.21
SUM+	**4.448**	**< .01**	**67**	**165**	**3.4**	**3.1**	**30**	**1.2**	**2.1**	**0.74**
TN	1.604	.11	3	13	2.9	2.1	1	1.0	1.7	0.89
TS	**2.459**	**< .01**	**31**	**35**	**1.7**	**1.8**	**10**	**.5**	**1.3**	**0.70**
Metacognitive monitoring										
FOK	1.404	.16	85	979	14.5	5.9	374	13.0	9.8	0.21
JOL	**2.528**	**.01**	**81**	**217**	**3.3**	**3.1**	**152**	**5.1**	**4.8**	**-0.46**
Test-taking strategies										
PDMI	**1.998**	**.05**	**77**	**295**	**4.4**	**2.6**	**107**	**3.6**	**4.4**	**0.25**
RRQ	**1.971**	**.05**	**84**	**966**	**14.9**	**8.1**	**569**	**17.7**	**12.9**	**-0.28**
RSE	1.652	1.00	84	1345	19.4	8.2	449	17.6	11.6	0.20
RSO	1.806	.07	73	273	4.0	3.0	150	6.0	7.6	-0.39
Vocabulary										
VOC-	1.300	.20	33	64	2.2	1.9	28	1.8	2.9	0.18
VOC+	**2.499**	**.01**	**73**	**152**	**2.5**	**1.8**	**39**	**3.1**	**11.9**	**-0.11**

Note: All codes are defined in Table 1. **Bold** indicates statistically significantly different (in the dependent-samples non-parametric *t* test evaluated at $p < .05$) between answer-correct and answer-incorrect questions

One elaboration code also showed differences; accurate knowledge elaboration from earlier in text (KEET+, *p* < .01, Cohen's *d* = 0.68), suggesting that examinees were encoding information from the passages as they answered the test questions; and when they accurately remembered information from an earlier passage, they were more likely to answer the question correctly. We note four differences with large (*d* ~ |0.50|) but non-significant effect sizes that might be suggestive for future research: Hypotheses (HYP) were verbalized 9 times by 6/86 participants and were more common when the question was answered correctly (*d* = 0.92). The other three effects were all more common when questions were answered incorrectly: incorrect global inferences (INFGLOB-, verbalized 64 times by 38/86 participants, *d* = -0.49), incorrect local inferences (INFLOC-, 83 times by 44/86 participants, *d* = -0.49), and incorrect knowledge elaborations from earlier in text (KEET-, 14 times by 12/86 participants, *d* = -0.47). All of these large-effect-size, non-significant effects are in the direction expected. Together, these findings are evidence in support of the explanation inference; correct IMRB responses result from the use of accurate inference-based strategies.

Second, we predicted that examinees should ***not*** differentially draw on accurate knowledge from previous coursework when answering questions correctly (vs. incorrectly). This hypothesis was supported for accurate prior knowledge activation from before the study (PKABS+, *p* = .52, Cohen's *d* = 0.12) and for accurate knowledge elaborations incorporating information from before the study (KEBS+, *p* = 1.00, Cohen's *d* = -0.23). Examinees did call on prior knowledge, but not differentially when answering questions correctly (vs. incorrectly).

Third, we predicted that examinees should ***not*** differentially draw on accurate vocabulary knowledge when answering questions correctly (vs. incorrectly). This hypothesis was supported, in that examinees answering questions incorrectly more often verbalized confidence in the meaning of a vocabulary word (VOC+, *p* = .01, Cohen's *d* = -0.11).

Fourth, we predicted that examinees should ***not*** differentially draw on test taking strategies when answering questions correctly (vs. incorrectly). This hypothesis was mostly supported, in that examinees verbalized two test-taking codes equally often when answering incorrectly as when answering correctly: process-of-elimination (RSE, *p* = 1.00, Cohen's *d* = 0.20) and other test-taking strategies (RSO, *p* = .07, Cohen's *d* = -0.39), and used one code—re-reading question—more often when answering incorrectly (RRQ, *p* = .05, Cohen's *d* = -0.28). One code, Passage Didn't Mention It (PDMI), was associated more with correct answers (*p* = .05, Cohen's *d* = 0.25).

Fifth, based on the general reading comprehension and study strategies literatures, we predicted that examinees might more often verbalize high-level strategies when answering correctly. This hypothesis was not supported, except that when answering correctly examinees more often made accurate summaries (SUM+; *p* < .01, Cohen's *d* = 0.74) and commented on the adequacy of the text (AT) when answering correctly but never when answering incorrectly (*p* = .04), but only by 5 of the 86 examinees. We note two differences with large (*d* ~ |0.50|) but non-significant effect sizes that might be suggestive for future research: Although non-significant and verbalized only 12 times by 10/86 participants, we note that statements about inadequacy of text (IT) were made more often when questions were answered correctly (*d* = 0.63). Also non-significant and verbalized only 14 times by 3/86 participants, we note that statements about taking notes (TN) were made more often when questions were answered correctly (*d* = 0.89).

Sixth, also based on the general reading comprehension and study strategies literatures, we predicted that examinees might less often verbalize low-level strategies when answering correctly. This hypothesis was not supported; two low-level codes were used *more* often when answering correctly—accurate paraphrase (PARA+; $p < .01$, Cohen's $d = 0.35$) and skimming (SKIM; $p = .02$, Cohen's $d = 0.06$). We note one difference with a large ($d \sim |0.50|$) but non-significant effect size that might be suggestive for future research. Although non-significant and verbalized only 24 times by 15/86 participants, we note that statements about importance of information (IOI) were made more often when questions were answered incorrectly ($d = -0.55$).

Seventh, and again based on the general reading comprehension and study strategies literatures, we predicted that examinees might more often verbalize feeling of knowing (FOK) when answering correctly and might less often verbalize judgment of learning (JOL) when answering incorrectly. The latter hypothesis was supported ($p = .01$, Cohen's $d = -0.46$).

These latter six findings are evidence almost entirely in support of the assumption that correct responses do not result from the use of other strategies (besides correct deductive inferences), with very limited exceptions; thus providing strong support for the explanation inference.

With regard to variability, examinees did in fact verbalize a larger variety of strategies when answering correctly ($M = 15.88$, $SD = 5.21$) than when answering incorrectly ($M = 9.81$, $SD = 5.33$, $t[85] = 8.07$, $p < .001$).

Discussion

Together, this study provides evidence in support of the description, explanation, and extrapolation inferences for the IMRB. In support of the description inference, the IMRB items were based on previously-collected think-aloud protocols from undergraduate students studying from an undergraduate biology text, with the same think-aloud passages used as stimuli in the IMRB, student-articulated accurate inferences from think-alouds used as correct item response options in the IMRB, and student-articulated inaccurate inferences from the think-alouds used as incorrect item responses in the IMRB. Thus, the passage and multiple-choice assessment tasks provide observable evidence of deductive inferences in biology, the relevant content area.

In support of the explanation inference, cognitive interviews showed that correct local and global inferences were made statistically significantly more often when questions were answered correctly versus incorrectly. By contrast, no statistically significant differences were seen on elaborative inferences, except when these involved learning from earlier in the text, which is a sign that examinees are encoding information from the texts while reading and answering questions.

Also in support of the explanation inference, correct responses were not, for the most part, associated with cognitive interviewing results on activation of background knowledge, use of other low- or high-level cognitive strategies, test-taking strategies, or vocabulary, all of which would represent sources of construct-irrelevant variance. Comparisons of verbalizations in correctly-answered versus incorrectly-answered questions were in support of the explanation inference for 4 of 5 background knowledge codes, 4 of 6 low-level cognitive strategy codes 7 of 9 testable high-level strategy codes, both metacognitive monitoring codes, 3 of 4 test-taking codes, and both vocabulary codes. Of the codes that were different between correctly-answered and incorrectly answered questions, two showed adaptive learning from previous IMRB passages (prior knowledge activation from earlier in text; PKAET+ and knowledge elaboration from earlier in text; KEET+) and one was an indication that the learner knew that he/she was not reasoning accurately (judgment of learning; JOL-). Of the remaining strategies showing differences, SKIM showed a very small effect size (*d* = 0.06), RRQ (re-read question) was used more when participants answered incorrectly, AT (adequacy of text) was used only 5 times and always when the question was answered correctly. Paraphrasing accurately (6.6% vs. 4.7%) and summarizing accurately (3.4% vs. 1.2%), use of text structure (1.7% vs. 0.5%), and the “Passage Didn’t Mention It” test-taking strategy (4.4% vs. 3.6%) do appear to advantage examinees, which, in context of the other findings, are minor evidence against the explanation inference.

Our cognitive interviewing evidence does not suggest that the IMRB is a test of background knowledge, in that verbalizations of neither accurate nor inaccurate prior knowledge differed significantly between correctly-answered and incorrectly-answered questions.

As predicted, low-level strategies in the cognitive interviews such as re-reading were not differentially associated with correctly-answered questions. Interestingly, accurate paraphrases (PARA; Cohen’s *d* = 0.35) and skimming (SKIM; Cohen’s *d* = 0.06) were associated with correctly-answered inference questions. Perhaps these surface strategies are used as means for advancing towards summarizing (SUM+), which is a high-level strategy that differed between correctly-answered and incorrectly-answered questions.

Summarizing (SUM+) was one of the most commonly verbalized high-level strategies and does appear to help students correctly answer inference-based questions. This is consistent with findings by Cromley and Wills (2016) that high-level strategies can be used as stepping stones to making inferences. We note also that only accurate paraphrases (PARA) and summaries (SUM) were associated with correctly-answered inferential questions, but not factually inaccurate applications of strategies.

Metacognitive monitoring was also frequently used; examinees’ sense that they were understanding or that they did have the correct answer (feeling of knowing; FOK) did not differentiate between correctly- and incorrectly-answered inference questions. Statements of *not* understanding (judgment of learning; JOL), were, however, more prevalent (Cohen’s *d* = -0.46) when examinees eventually gave an incorrect answer to the inference question. Thus, these statements appear to be a sign of accurate monitoring, but with inadequate fixup strategies (i.e., ones enabling the examinee to eventually come to a correct conclusion).

Interestingly, verbalizations about understanding vocabulary were statistically significantly more common (Cohen's *d* = -0.11) when examinees eventually gave an incorrect answer to the inference-based question. Perhaps examinees did know the meaning of the vocabulary words but were not able to build from the meaning of the words to an integrated understanding of the passage.

The cognitive interviewing results suggest the IMRB is not a measure of test-taking skills. Although examinees seemed to drop immediately into a test-taking "set" as noted by Rupp et al. (2006) and verbalized test-taking strategies more than 4000 times, re-reading the question was in fact statistically significantly associated with incorrectly-answered inference questions. Thus, we might interpret re-reading the question as an indication of lack of comprehension; and thereby, makes the action of producing an accurate inference unlikely. The only test-taking strategy that did seem to statistically significantly help was stating that "the passage didn't mention it". We would not advocate for teaching this strategy (as it misled examinees in 107 instances), but overall, it does seem to be more associated with correctly-answered inference-based questions.

Our finding of a larger variety of moves (strategies, inference, monitoring, and so on) when examinees gave correct answers is consistent with a small literature in reading comprehension. Specifically, "good" comprehenders adapt in real-time to the challenges of the text(s) they are reading and use strategies in a flexibly adaptive way (Cho & Afflerbach, 2017), whereas "poor" comprehenders seem to get locked into inflexible, habitual, maladaptive patterns

(Cromley & Wills, 2016). If what we see from these examinees is the same pattern, it suggests that the flexibly adaptive strategy user model can be generalized from text comprehension to include assessments of inference, at least for undergraduate biology students.

In support of the extrapolation inference, the stimuli and questions that were designed to elicit inferences did indeed elicit inferences (*N* = 978 inferences), and 574 of these (7.5% of all verbalizations) were the targeted accurate local and global inferences. The use of passages from the previous think-alouds, together with questions designed to elicit the correct inferences, and distractors drawing on incorrect inferences that had been verbalized in the think-alouds, did indeed lead the participants in the present study to draw inferences from the presented information. This supports the inference from classification of examinees based on IMRB scores to their level of deductive reasoning skill using newly-presented information (i.e., the extrapolation inference).

In summary, our cognitive interviews provide strong evidence for the explanation and extrapolation inferences in the interpretive argument for the IMRB. Furthermore, our item development procedure provides some support for the description inference. Users can be confident that observed scores on the IMRB represent an examinee’s skill at deductive reasoning with new biology content, of the type that undergraduate biology students need to engage in to understand what they are learning. Other validation work by our team supports the use of the instrument for student placement into courses (for advisor and instructor use) and to predict course grades (for researcher use).

References

Ahmed, Y., Francis, D. J., York, M., Fletcher, J. M., Barnes, M., & Kulesz, P. (2016). Validation of the direct and inferential mediation (DIME) model of reading comprehension in grades 7 through 12. *Contemporary Educational Psychology*, *44*, 68-82.

Alexander, P. A., & Kulikowich, J. M. (1991). Domain knowledge and analogic reasoning ability as predictors of expository text comprehension. *Journal of Reading Behavior, 23*(2), 165-190.

Ash, G. E., & Baumann, J. F. (2017). Vocabulary and reading comprehension: The nexus of meaning (pp. 377-405). In S. E. Israel (Ed.), *Handbook of research on reading comprehension, second edition.* NY, NY: Routledge.

Bonner, J., & Holliday, W. (2006). How college science students engage in note-taking strategies. *Journal of Research in Science Teaching, 43*(8), 786-818.

Cain, K., & Oakhill, J. (2014). Reading comprehension and vocabulary: Is vocabulary more important for some aspects of comprehension? *L'Année Psychologique*, *114*(4), 647-662.

Cho, B.-Y., & Afflerbach, P. (2017). An evolving perspective of constructively responsive reading comprehension strategies in multilayered digital text environments (pp. 109-134). In S. E. Israel (Ed.), *Handbook of research on reading comprehension, second edition.* NY, NY: Routledge.

Cromley, J. G. & Azevedo, R. (2007). Testing and refining the direct and inferential mediation model of reading comprehension. *Journal of Educational Psychology, 99*(2), 311-325. doi: 10.1037/0022-0663.99.2.311

Cromley, J. G., Snyder-Hogan, L. E., & Luciw-Dubas, U. A. (2010a). Cognitive activities in complex science text and diagrams. *Contemporary Educational Psychology, 35*, 59–74. doi: 10.1016/j.cedpsych.2009.10.002

Cromley, J. G., Snyder-Hogan, L. E., & Luciw-Dubas, U. A. (2010b). Reading comprehension of scientific text: A domain-specific test of the direct and inferential mediation model of reading comprehension. *Journal of Educational Psychology, 102*(3), 687-700. doi: 10.1037/a0019452

Cromley, J., & Wills, T. W. (2016). Flexible strategy use by readers who learn much versus learn little: Transitions within think-aloud protocols. *Journal of Research in Reading, 39*(1), 50–71. doi:10.1111/1467-9817.12026

Dai, T., & Cromley, J. G. (2014). Changes in implicit theories of ability in biology and dropout from STEM majors: A latent growth curve approach. *Contemporary Educational Psychology*, *39*(3), 233-247. doi: 10.1016/j.cedpsych.2014.06.003

Daugaard, H. T., Cain, K., & Elbro, C. (2017). From words to text: inference making mediates the role of vocabulary in children's reading comprehension. *Reading and Writing*, *30*(8), 1773-1788.

Dole, J. A., Valencia, S. W., Greer, E. A., & Wardrop, J. L. (1991). Effects of two types of prereading instruction on the comprehension of narrative and expository text. *Reading Research Quarterly, 26*(2), 142-159.

Ehrlich, M. F., Remond, M., & Tardieu, H. (1999). Processing of anaphoric devices in young skilled and less skilled comprehenders: Differences in metacognitive monitoring. *Reading and Writing*, *11*(1), 29-63.

Farr, R., Pritchard, R., & Smitten, B. (1990). A description of what happens when an examinee takes a multiple-choice reading comprehension test. *Journal of Educational Measurement, 27*(3), 209-226.

Flippo, R. F., & Caverly, D. (2009). *Handbook of college reading and writing.* Mahwah, NJ: Erlbaum.

Gilabert, R., Martínez, G., & Vidal-Abarca, E. (2005). Some good texts are always better: Text revision to foster inferences of readers with high and low prior background knowledge. *Learning and Instruction*, *15*(1), 45-68.

Heist, B. S., Gonzalo, J. D., Durning, S., Torre, D., & Elnicki, D. M. (2014). Exploring clinical reasoning strategies and test-taking behaviors during clinical vignette style multiple-choice examinations: a mixed methods study. *Journal of Graduate Medical Education*, *6*(4), 709-714.

Hernandez, P. R., Schultz, P. W., Estrada, M., Woodcock, A., & Chance, R. C. (2013). Sustaining optimal motivation: A longitudinal analysis of interventions to broaden participation of underrepresented students in STEM. *Journal of Educational Psychology, 105*(1), 89-107.

Hong, E., Sas, M., & Sas, J. C. (2006). Test-taking strategies of high and low mathematics achievers. *The Journal of Educational Research, 99*(3), 144-155. http://www.jstor.org/stable/27548124

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1-73.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge, England: Cambridge University Press.

Koskey, K. L., Karabenick, S. A., Woolley, M. E., Bonney, C. R., & Dever, B. V. (2010). Cognitive validity of students' self-reports of classroom mastery goal structure: What students are thinking and why it matters. *Contemporary Educational Psychology*, *35*(4), 254-263.

Lawson, A. E., Banks, D. L., & Logvin, M. (2007). Self-efficacy, reasoning ability and achievement in college biology. *Journal of Research in Science Teaching, 44*(5), 706-724.

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

Maltese, A. V., & Tai, R. H. (2011). Pipeline persistence: Examining the association of educational experiences with earned degrees in STEM among US students. *Science Education*, *95*(5), 877-907.

Martin, J. H., Montgomery, R. L., & Saphian, D. (2006). Personality, achievement test scores, and high school percentile as predictors of academic performance across four years of coursework. *Journal of Research in Personality, 40*(4), 424-431.

McNamara, D. S. (2004). SERT: Self-explanation reading training. *Discourse Processes, 38*(1), 1-30, doi: 10.1207/s15326950dp3801_1

McNamara, D. S., Levinstein, I. B., & Boonthum, C. (2004). iSTART: Interactive strategy training for active reading and thinking. *Behavior Research Methods, Instruments, & Computers*, *36*(2), 222-233.

Otero, J., León, J. A., & Graesser, A. C. (Eds.). (2002). *The psychology of science text comprehension*. Mahwah, NJ: Lawrence Erlbaum Associates.

Ouellette, G. P. (2006). What's meaning got to do with it: The role of vocabulary in word reading and reading comprehension. *Journal of Educational Psychology*, *98*(3), 554-566.

Ozuru, Y., Dempsey, K., & McNamara, D. S. (2009). Prior knowledge, reading skill, and text cohesion in the comprehension of science texts. *Learning and Instruction*, *19*(3), 228-242.

Perfetti, C. (2007). Reading ability: Lexical quality to comprehension. *Scientific Studies of Reading*, *11*(4), 357-383.

Piaget, J. (1952). *The origins of intelligence in children* (M. Cook, Trans.). NY, NY: International Universities Press.

Reich, G. A. (2009). Testing historical knowledge: standards, multiple-choice questions and student reasoning. *Theory & Research in Social Education, 37*(3), 325-360. doi: 10.1080/00933104.2009.10473401

Rupp, A. A., Ferne, T., & Choi, H. (2006). How assessing reading comprehension with multiple-choice questions shapes the construct: A cognitive processing perspective. *Language Testing*, *23*(4), 441-474.

Segers, E., & Verhoeven, L. (2016). How logical reasoning mediates the relation between lexical quality and reading comprehension. *Reading and Writing*, *29*(4), 577-590.

Smith, M. D. (2017). Cognitive validity: Can multiple-choice items tap historical thinking processes? *American Educational Research Journal*, *54*(6), 1256-1287.

Surber, J. R., & Schroeder, M. (2007). Effect of prior domain knowledge and headings on processing of informative text. *Contemporary Educational Psychology*, *32*(3), 485-498.

Taboada, A., & Guthrie, J. T. (2006). Contributions of student questioning and prior knowledge to construction of knowledge from reading information text. *Journal of Literacy Research*, *38*(1), 1-35.

Van den Broek, P. (2010). Using texts in science education: Cognitive processes and knowledge representation. *Science*, *328*(5977), 453-456.

Van Den Broek, P., & Kendeou, P. (2008). Cognitive processes in comprehension of science texts: the role of co-activation in confronting misconceptions. *Applied Cognitive Psychology*, *22*(3), 335-351.

Van Den Broek, P., Virtue, S., Everson, M. G., Tzeng, Y., & Sung, Y. C. (2002). Comprehension and memory of science texts: Inferential processes and the construction of a mental representation (pp. 131-154). In J. Otero, J. A. Leon, & A. C. Graesser (Eds.), *The psychology of science text comprehension.* Mahwah, NJ: Erlbaum.

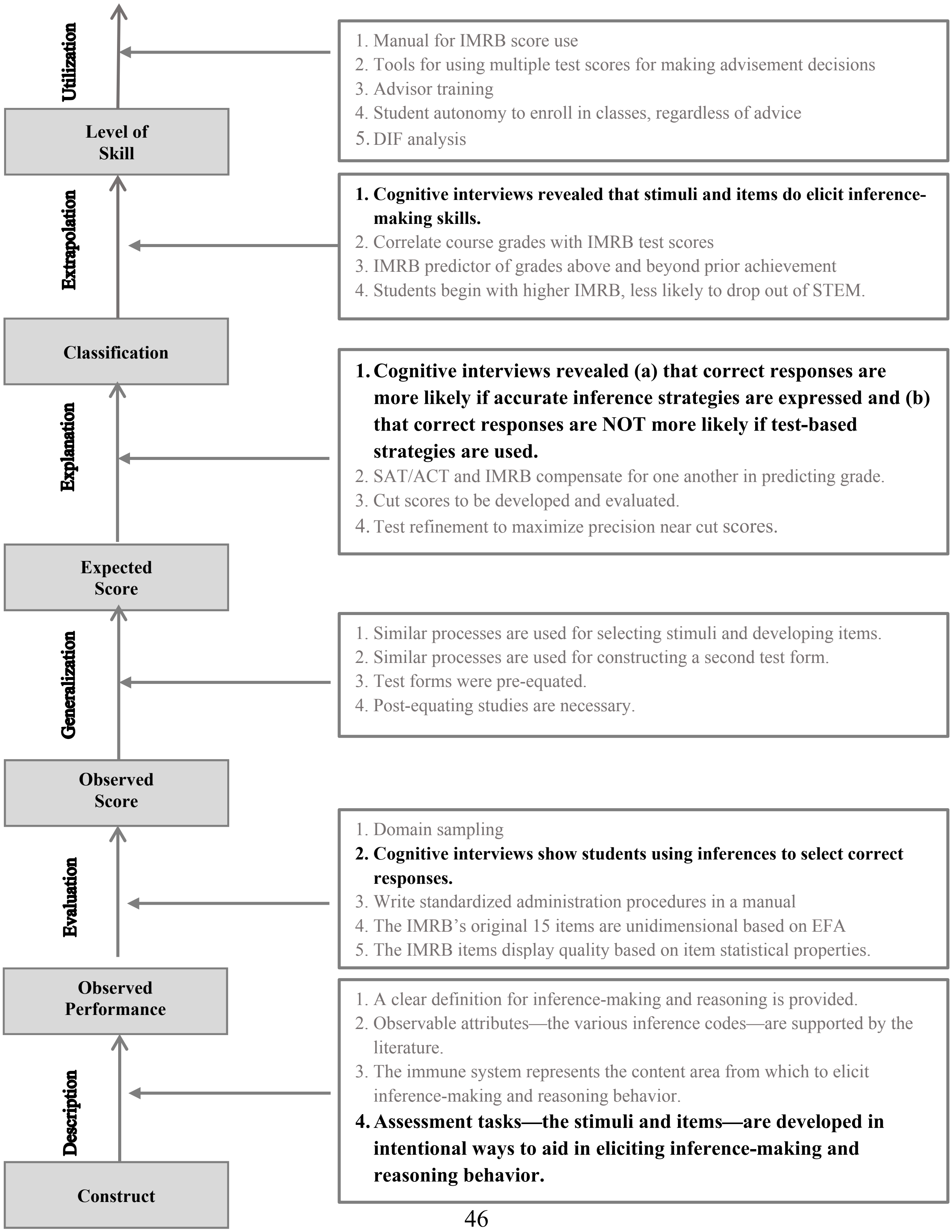
Willis, G. (2018). Cognitive interviewing in survey design: State of the science and future directions. In D. L. Vannette & J. A. Krosnick (Eds.), *The Palgrave handbook of survey research* (pp. 103-107). Cham, Switzerland: Palgrave Macmillan.

Willson, S., & Miller, K. (2014). Data collection. In K. Miller, S. Willson, V. Chepp, & J.-L. Padilla (Eds.), *Cognitive interviewing methodology.* NY, NY: John Wiley & Sons.

Zusho, A., Pintrich, P. R., & Coppola, B. (2003). Skill and will: The role of motivation and cognition in the learning of college chemistry. *International Journal of Science Education, 25*(9), 1081-1094.

Appendix *Interpretive Argument for IMRB* This research in **bold**, other studies grey

Conclusion: IMRB test scores are useful for providing information to university personnel for advising students on biology courses best suited for them and for identifying students who may benefit from supports that would increase reasoning skills and thereby encourage retention in STEM majors.



Utilization

1. Manual for IMRB score use
2. Tools for using multiple test scores for making advisement decisions
3. Advisor training
4. Student autonomy to enroll in classes, regardless of advice
5. DIF analysis

Level of Skill

Extrapolation

1. **Cognitive interviews revealed that stimuli and items do elicit inference-making skills.**
2. Correlate course grades with IMRB test scores
3. IMRB predictor of grades above and beyond prior achievement
4. Students begin with higher IMRB, less likely to drop out of STEM.

Classification

Explanation

1. **Cognitive interviews revealed (a) that correct responses are more likely if accurate inference strategies are expressed and (b) that correct responses are NOT more likely if test-based strategies are used.**
2. SAT/ACT and IMRB compensate for one another in predicting grade.
3. Cut scores to be developed and evaluated.
4. Test refinement to maximize precision near cut scores.

Expected Score

Generalization

1. Similar processes are used for selecting stimuli and developing items.
2. Similar processes are used for constructing a second test form.
3. Test forms were pre-equated.
4. Post-equating studies are necessary.

Observed Score

Evaluation

1. Domain sampling
2. **Cognitive interviews show students using inferences to select correct responses.**
3. Write standardized administration procedures in a manual
4. The IMRB's original 15 items are unidimensional based on EFA
5. The IMRB items display quality based on item statistical properties.

Observed Performance

Description

1. A clear definition for inference-making and reasoning is provided.
2. Observable attributes—the various inference codes—are supported by the literature.
3. The immune system represents the content area from which to elicit inference-making and reasoning behavior.
4. **Assessment tasks—the stimuli and items—are developed in intentional ways to aid in eliciting inference-making and reasoning behavior.**

Construct