DO RUBRICS IMPROVE STUDENTS’ METACOMPREHENSION ACCURACY?

A thesis submitted
To Kent State University in partial
Fulfillment of the requirements for the
Degree of Master of Arts

By
Christina Marie Poulin

August 2013
Thesis written by
Christina M. Poulin
B.A., Albion College, 2009
M.A., Kent State University, 2013

Approved by
___John T. Dunlosky, Ph.D________, Advisor
___Maria Zaragoza, Ph.D___________, Chair, Department of Psychology
___James L. Blank, Ph.D_____________, Dean, College of Arts and Sciences

ii
TABLE OF CONTENTS

LIST OF FIGURES………………………………………………………………..iv
INTRODUCTION…………………………………………………………………1
EXPERIMENT 1…………………………………………………………………..9
    Method……………………………………………………………………..9
    Participants………………………………………………………………9
    Materials………………………………………………………………..9
    Procedure………………………………………………………………10
    Results…………………………………………………………………12
EXPERIMENT 2…………………………………………………………………..15
    Method…………………………………………………………………….15
    Results…………………………………………………………………16
GENERAL DISCUSSION………………………………………………………….18
REFERENCES………………………………………………………………………22
TABLES……………………………………………………………………………..25
FIGURES…………………………………………………………………………….27
APPENDICES………………………………………………………………………32
LIST OF FIGURES

FIGURE 1. Mean relative accuracy as a function of group and question type in Experiment 1. Error bars report standard error of the mean.

FIGURE 2. Mean absolute accuracy bias scores as a function of group in Experiment 1. Error bars report standard error of the mean.

FIGURE 3. Mean relative accuracy as a function of group and question type. Error bars report standard error of the mean.

FIGURE 4. Mean absolute accuracy bias score as a function of group for Experiment 2. Error bars report standard error of the mean.
INTRODUCTION

In school settings, students are often assigned to read short texts across different content areas. If a student were assigned a text about avalanches to learn, a student may gauge how well they understood the content in the text prior to being given a test. Students could benefit from making judgments during text learning, but only if those judgments are accurate. Judgment accuracy is comprised of two kinds of accuracy: relative accuracy and absolute accuracy. Relative accuracy refers to the ability for students to discriminate between their performance on one test over another one and is defined by the correlation of their judgment made during study and their final test grade. Absolute accuracy differs from relative accuracy in that it relates to how their perceptions of how well they think they learned the content matches their overall test performance. Metacomprehension research has consistently shown that students have poor judgment accuracy. Thus, the main objective of the studies was to develop a tool based off of rubrics commonly used in education, and to observe how students’ judgment accuracy of future test performance from reading expository texts changes with the inclusion of this metacognitive tool. Before I describe my current approach, I will discuss previous approaches to improving judgment accuracy and each of their costs and benefits to having students engage with these evaluation methods.
Prior strategies to improve judgment accuracy have yielded some improvement in students’ judgment accuracy. One such tool that has been used to potentially increase judgment accuracy for text comprehension is rereading. Consider the following evidence from a set of studies that included this strategy. Rawson, Dunlosky, and Thiede (2000) had students read a text either once or twice because rereading encourages people to process a text for a further deep understanding. They observed that rereading did increase relative accuracy more than after a single reading of an educational text. Rereading has also been thought to promote surface as well as text levels of representation, and that further reading helps to build a situation model (Wiley, Griffin, & Thiede, 2003). A situation model, briefly speaking, is the mental representation of what a text is about compiled during and after studying a text. However, studies have not conclusively supported the efficacy of rereading for text comprehension (Rawson & Dunlosky, 2005). By reading a text more than once, little change is made to the situational model (Callender & McDaniel, 2009; Mills et al., 1998). When the rereading takes place after the initial reading it can have little effect on building the situation model. For example, Dunlosky, Rawson, & Middleton (2005) found that rereading immediately after studying Graduate Requirement Exam sample texts increased judgment accuracy and had no significant change at a delayed re-reading. Rereading might also give students the false perception that they comprehend educational texts (Wiley, Griffin, & Thiede, 2005). When students engage with the text a second time, it may enhance fluency and readers might be more likely to confuse this fluent experience with a firm understanding of the text. Rereading might be beneficial for getting a surface level understanding of
educational texts but it is not clear why rereading could enhance judgment accuracy entirely on its own.

Another possible study strategy selected to improve metacomprehension accuracy is summarization. Summarization can help students organize and identify main ideas in educational texts (Bretzing & Kulhavy, 1979). Studies have also demonstrated that summarization allows students to engage in more active processing of the text when they summarize it compared to passive reading (Dunlosky et al, 2013). For example, Thiede and Anderson (2003) had students read expository texts over a variety of topics. In one condition, students were told to make a metacomprehension judgment for each text while the other condition was asked to generate a summary of the text prior to completing the judgment. They observed that the students who generated the summaries had higher judgment accuracy than when they did not. However, using summarization to improve accuracy has its limitations. The quality of summaries generated by students also matters to determine how well they understood the texts. If students cannot correctly identify what is considered a “strong” summary, they might not get the most out of using a summarization strategy. Prior knowledge of the material that should be understood for a future test can also impact the quality of generated summaries. If these students are producing poorly constructed summaries then it might not enhance their judgment accuracy much for predicting subsequent test performance. While summarization might be beneficial to structuring the content during study, it overlooks the research that has consistently shown that students are highly overconfident in their future academic success prior to and after reading educational materials (Dunlosky & Lipko, 2007;
Hacker, Dunlosky, & Graesser, 1998; Thiede, 2009). To foreshadow the later discussion of summarization’s costs, it might not be practical or feasible as a sole strategy to improve students’ judgment accuracy for educational texts because it is time-consuming and might not work well for students with low knowledge.

Another strategy that has been implemented to improve metacomprehension accuracy is keyword generation. The structure of a keyword generation task goes as follows: students are told to generate keywords following reading a text (Thiede et al., 2005; Thiede, 2009). The studies followed a similar paradigm; some students were given varied temporal delays; either an immediate or delayed keyword generation test. Students would attempt to generate keywords that best represented the ideas in the texts they previously read. Those students who had the delayed keyword generation condition had higher judgment accuracy compared to those students who had either no keyword generation or immediate keyword generation. By implementing the keyword generation task, it allows students to further consider the ideas in the text. This is especially true if the keyword generation task includes a delayed test (Thiede et al., 2003; Thiede et al., 2005).

Cost and Benefits of Previously Implemented Strategies

While the discussed strategies for improving metacomprehension accuracy have been beneficial in some educational contexts, they also have been met with some related costs in terms of feasibility and generalizability to all students. Rereading is not really effective for promoting greater understanding of texts and may give students the illusion of fluency that leads to overconfidence (Wiley et al., 2005). It might also constrain the
ability to actively work with the text and cause some rigidity in terms of their understanding.

Summarization also comes with its own downfalls. If students are not creating quality summaries, it would defeat the purpose of the summarization task. This is especially true if students have poor comprehension of the texts. Utilizing summarization as a sole strategy for increasing judgment accuracy could also take away time from other study activities that might be more effective such as self-explanation, or putting ideas in the text into their own words. To complicate the implementation of this strategy, Dunlosky et al. (2012) prescribed summarization as a low utility study strategy, meaning that it does not provide the most effective gains for long-term retention of educational materials. It also has limited scope for implementation in an educational context because continually generating summaries for all subject areas would take up more effort than necessary with little to no gains in academic success in the long-term.

Keyword generation is the most promising strategy of the three described applications for improving metacomprehension accuracy. However its effectiveness is somewhat limited because it does not entirely alleviate students’ overconfidence of how much they understood from texts (Lipko & Dunlosky, 2007; Thiede, 2010). Thiede et al. (2005) discussed that more comparisons with other potential strategies for improving metacomprehension should be welcome so it does open possibilities for developing new applications.
Rubrics

The goal of these studies was to develop and apply an easy-to-use technique that may help enhance students’ metacomprehension accuracy. A rubric, loosely defined, is an evaluation tool used to assess one’s content or understanding of educational materials by separating content into sub-categories and is defined by numerical values. In education, rubrics have been used as an additional form of grading to gauge performance on class assignments. Rubrics might be an excellent candidate to improve metacomprehension accuracy because they would allow students to focus on a text’s main ideas while also building a situation model that further develops a greater understanding of the text.

Education researchers have specified a few rubric formats that have been used frequently in course curriculum. Rubric formats generally fit into one of two types: holistic or analytical (Allen & Tanner, 2006). Holistic rubrics take together the ideas of the text or assignment and collapse them into general ideas to consider for inclusion in the answers. For example, in order to receive a full credit for a test question, a student might evaluate himself or herself via a question such as “Do you have a clear understanding of the material represented in this chapter?” Analytical rubrics identify specific parameters for topics that should be included in the assignment. This might include several subcategories with detailed information that will need to be included in order to receive certain amounts of credit. Despite the overwhelming use of rubrics in education, there is limited evidence for their effectiveness (Andrade, 2000; Andrade,
2001; Andrade, H., 2006; Andrade, Du, & Wang, 2008), and to date, no one has tested rubrics in a controlled laboratory setting.

For this set of experiments, rubrics were developed that could be adapted to a variety of text materials. Two rubric formats were created based upon the rubric categories defined by Allen and Tanner (2006): general and specific rubrics (see Appendix B). The general rubric is similar to the holistic rubric in that it states generalities about how students should evaluate themselves on how well they understood that text segment, regardless of the content shown. This format stayed static across all of the text segments presented. The specific rubric was adapted from the analytical rubric design. It included information from each text segment and told students to compare their understanding of the material to a grade of how much information they understood. In addition, we wanted to look at having students focus on the rubric for a certain period of time and how that would affect their accuracy.

Rubrics are a strong candidate for improving metacomprehension accuracy because they will direct students to cues in the text that will be important for them to understand for later testing. They may also further slow down students’ reading process to make sure they can check their learning progress without costing them too much time or effort. I predict that students that receive rubrics during study will have better metacomprehension accuracy than students who receive no rubric. Also, students that receive the specific rubrics during study are expected to have better metacomprehension accuracy than students with the general rubric because it is a more fine-grained format.
that allows them to focus on the detailed information that they need to know well for the final test.

**Current Experiments**

In order to test the efficacy of rubrics as an tool to improve metacomprehension accuracy, this set of experiments followed this structure: initial study, judgment evaluation, and final test. First, students were instructed to read a set of expository texts. Some students would have different directions based on their randomly assigned condition (rubric or no rubric). Following reading through each of the texts, students were asked to judge how well they understood the text by each segment. The judgment evaluation would include the rubric (Fig. 1) based on their assigned condition. Students would make their metacomprehension judgment before proceeding to the next text segment. When they were finished making judgments for each of the texts, a final recall test was given. The final recall test asked students both detailed and conceptual questions for each of the texts.
Experiment 1

Method

Participants. Ninety four undergraduates from Kent State University participated in this study to partially fulfill an Introductory Psychology requirement. Participants were randomly assigned to one of three groups: no rubric ($n = 32$), general rubric ($n = 31$), or specific rubric ($n = 31$).

Materials. The two expository texts were the same as those used in Thomas and McDaniel (2007). The texts, “Kanchenjunga Mountain,” and “Spiders” were between 543 and 685 words long. Each text was split into 6 smaller text segments that ranged from 60 to 120 words long. There were 24 final test questions: 12 detailed and 12 conceptual questions that were designed to have students focus on their comprehension of the texts. Each text had 6 of each final question category: detailed and conceptual. The detailed questions were designed to focus attention to the specific details included in the texts. For example, if a participant was asked to fill in the blank in the following recall statement, “Mountains of Kanchenjunga are considered dangerous not only because of the ice avalanches but also because of the _______________,” they would be engaging with the specific details from the text segment. A conceptual question would be asking “Why are avalanches not as large in the Alps compared to those in the Kanchenjunga
mountains?” It addresses the participant’s ability to connect different ideas from the text segment into a cohesive statement and addresses their comprehension of the text as a whole. Some of these questions were identical to the original materials used in the Thomas and McDaniel (2007) experiments, but 2 of the detailed questions and 6 of the conceptual questions were modified. The main reasoning behind modifying some of the questions was because several of them did not attempt to address the specific processes, such as the original conceptual questions not focusing on the overall understanding of the text. To look at further question examples, please reference Appendix A.

For the rubric design, I adapted the rubric formats suggested by Allen and Tanner (2006). Both rubrics contained values as well as descriptions of what would constitute as a measure of comprehension for that particular value. The following values were placed above each of the description containers: no credit, half credit, and full credit (Appendix B). The rubric container descriptions differed in content based on whether it was a general or specific rubric. General rubric description containers gave information to examine their overall understanding of the material. The general rubric stayed the same for all texts presented and across each text segment. Specific rubric containers placed emphasis on important content from each of the presented text segments that the reader should consider in their understanding of the text. These descriptors would change for each of the text segments so the student would be evaluating their understanding of specific content.

Procedure. Participants worked individually at computers. Following reading through and signing for informed consent, participants were told to read instructions
carefully, sit quietly, and take as much time as they needed on the task. Participants began the study by reading the instructions about the reading and rubric evaluation tasks. Based on their assigned condition, all students were told that they would be receiving some kind of judgment evaluation format following reading each text: standard comprehension judgment, general rubric, or specific rubric evaluation. An example of the rubric format was told to the participants in the initial instructions. Following the initial instructions, the first of the two texts was presented. The order of the text presentation was counterbalanced so half of the students would receive the “Avalanche” text first and then the “Spider” text next. During the reading portion of the study, the text was presented into much shorter segments that participants could proceed through at their own pace. After reading the first text, participants made a metacomprehension judgment, with or without the accompanying rubric based on condition assignment. For participants in one of the two rubric conditions (general or specific), a rubric would be given along with its matched text segment in the order that the participant read through them. At this time, they would be asked to evaluate against the rubric on a scale of 1 to 100 how well they thought they were going to perform on questions that would test their understanding of the concepts within that text segment. Once they had made their metacomprehension judgment, they would continue to make judgments for the next 5 text segments with the rubric feedback. For the participants in the no rubric group, they were instructed to make a metacomprehension judgment based on the text segment presented to them. After finishing the judgments for the first text, the participant would be given the second text to read and would proceed through the same reading and metacomprehension judgment
evaluation with the accompanying rubric or without one. When the second text had been studied and evaluated, the final test was given to the participant. They were asked to answer 12 questions for each text which targeted their conceptual understanding and specific details. In a space provided below each question, participants typed in their answers. The answers ranged from single words to complete sentences. Participants could take their time answering each of the final test questions. When they were finished completing the final test, participants received a brief feedback sheet about the experiment.

Results

Metacomprehension Ratings and Performance. To revisit, metacomprehension accuracy is the relationship between metacomprehension judgments and performance on the final recall test of reading comprehension. In order to understand how the metacomprehension judgments and performance relate to one another, descriptive statistics will be reported first. The median proportion of correct recall test responses and metacomprehension judgments across the two texts was computed for each participant. The reason for choosing to report the median is because it is the best measure of central tendency when there is a small set of scores where extreme scores could affect the mean (Gravetter & Wallanau, 1999). The mean of the medians was then computed across participants in each rubric condition. A one-way ANOVA revealed that recall for detailed questions was not significant across conditions, $F(2,102) = .170$, $MSE = .007$, $p = .844$, $\eta^2_p = .001$. Recall for conceptual questions was also not significant across the rubric conditions, $F(2,102) = .671$, $MSE = .024$, $p = .513$, $\eta^2_p = .001$ (see Figure 1). The main effect of rubric group on metacomprehension judgments were found to be significant across conditions,
Bonferroni post-hoc analyses showed that metacomprehension judgments (see Table 1, right hand column) showed that metacomprehension judgments were higher in the specific rubric condition ($M=76.29, SD=18.68$) than in the general rubric condition ($M=63.92, SD=17.67$); no other differences were significant.

**Metacomprehension Accuracy.** The most commonly used measure of metacomprehension accuracy is the Goodman-Kruskal gamma correlation, which takes the metacomprehension judgments and performance on the final recall test computed across texts. Two gamma correlations were computed for each student; one between judgments and detailed question accuracy and another between judgments and conceptual question accuracy. The mean intra-individual correlation was then computed across participants for each condition. Seven participants had indeterminate correlations due to a lack of variance in their judgments. One-way ANOVAs showed that metacomprehension accuracy did not differ across conditions for either the detailed questions, $F(2,95) = 1.361, MSE=.277, p=.261, \eta^2_p=.01$, or the conceptual questions, $F(2,96) = 1.062, MSE = .221, p = .350, \eta^2_p=.01$ (see Figure 1). While we did replicate that having no metacognitive tool would lead to poor metacomprehension accuracy, the general rubric reaching at only produced an accuracy level of .35, which does not reflect similar results from using other metacognitive tools such as the delayed-keyword generation task (Thiede, 2003).

Absolute accuracy is another measure of metacognitive accuracy that examines the precision of the judgments with the overall test performance. One way to measure absolute accuracy is by computing the absolute accuracy bias index. The bias index indicates when a participant is over- or underconfident when making a metacomprehension judgment. Absolute accuracy bias is calculated by taking the mean performance total and subtracting
from it the confidence judgment magnitude mean times 100. A one-way ANOVA demonstrated no differences between conditions on total absolute accuracy, $F(2,102) = 1.927$, $MSE = 1012.72$, $p = .211$, $\eta^2_p = .01$. Absolute accuracy bias for detailed questions was trending towards significance but did not reach the criterion alpha < .05, $F(2,102) = 1.192$, $MSE = 874.72$, $p < .15$, $\eta^2_p = .01$.

In sum, participants were more overconfident when given the specific rubric than when they were given a general rubric that assessed their overall knowledge of the texts. In addition, a trend showed that metacomprehension accuracy faired better with the general rubric condition in comparison to the specific rubric and no rubric conditions. However, it seemed that subjects were not utilizing the general rubric effectively in a self-paced setting and not even considering it when they made their metacomprehension judgments. Experiment 2 was designed to further explore the use of rubrics and how it might be used effectively.
Experiment 2

Experiment 2 was designed to overcome one concern of the rubric design used in Experiment 1. During the rubric and judgment evaluation phase of Experiment 1, participants were told to examine the rubric and the text segment to make a metacomprehension judgment at their own pace. However, the time participants spent looking at the rubrics was quite short in comparison to the no rubric group. Because the rubric is meant to be used as an evaluation tool for the students during the judgment phase, it was concerning that the participants were speeding through the judgment phase and perhaps were not even looking at the rubric. To encourage use of the rubric (and in turn improve metacomprehension accuracy), a timed rubric group was used for this experiment; namely, participants in the timed rubric group would be given an allotted time to study with the rubric before proceeding to make their metacomprehension judgment. This rubric condition replaced the “specific” rubric used in Experiment 1 that resulted in much higher judgments but lower accuracy. Other than the changes to the rubric conditions, the texts and comprehension test remained identical to the materials in Experiment 1.

Method

One hundred six undergraduate students participated for course research credits. The materials, design, and procedure were identical to Experiment 1 except for the changes made to the rubric groups. For Experiment 2, the specific rubric group from Experiment 1
was changed to the timed rubric group. In the timed rubric group, the general rubric would be shown with the accompanying text segment for ten seconds before the metacomprension judgment rating box would be shown. The participants would not be allowed to make a judgment until the time was up.

Results

Metacomprension Ratings and Performance. To revisit, metacomprension judgments and performance scores comprise relative accuracy scores so these descriptives will be reported first (see Table 2). A one-way ANOVA showed that there were no significant differences in final test performance, $F(2,103) = 1.730$, $MSE = .048$, $p<.182$, $\eta_p^2=.002$. Furthermore, there were no significant differences found in the specific test question performance. However, metacomprension judgment magnitude was also not found to be significant, $F(2,103)=.624$, $MSE = 210.70$, $p<.538$, $\eta_p^2=.004$.

Metacomprension Accuracy. For relative accuracy, two gamma correlations were computed for each participant (one for detailed and one for conceptual question accuracy, as in Experiment 1). These were then computed across all participants to create a mean intra-individual correlation across participants was then computed for each of the rubric groups (see Figure 3). Nine participants were omitted because of a lack of variance in their metacomprension judgments. One-way ANOVAs showed that metacomprension accuracy did not differ across groups for either the detailed questions, $F(2,97) = .286$, $MSE=.066$, $p=.752$, $\eta_p^2=.003$, or the conceptual questions, $F(2,100) = .177$, $MSE = .048$, $p =.838$, $\eta_p^2=.002$. No further differences were significant.

For absolute accuracy (Figure 2), a one-way ANOVA showed no significant differences between the rubric conditions, $F(2,103) = 1.471$, $MSE = 3099.61$, $p=.235$, $\eta_p^2=$
.04 (see Figure 4). Although the general rubric condition is showing a trend towards less bias than the other two conditions, the difference is not statistically significant.

Overall, Experiment 2 demonstrated that the addition of the timed rubric did not significantly change students' use of the rubric. While the general rubric still gains some promise by decreasing overconfidence, it is still not showing how it could be more effective than studying without it.
GENERAL DISCUSSION

Metacomprehension research has established that students are inaccurate at predicting their future test performance from studying text materials, with more recent research showing that using metacognitive tools such as rereading and delayed keyword generation can improve students’ metacomprehension accuracy (Lipko & Dunlosky, 2007; Thiede, 2003). In the present experiments, rubrics were explored to evaluate their effect on metacomprehension accuracy, with emphasis on how it would improve relative and absolute accuracy. Across the two experiments, relative accuracy was consistently poor for students, demonstrating how difficult it is to improve metacomprehension accuracy, even when metacognitive tools are introduced. In Experiment 1, relative accuracy was shown to be no different across the different rubric conditions, despite a trend for rubric conditions’ accuracy being better than being given no rubric at all. In Experiment 2, the addition of a timed rubric also did not demonstrate differences between the rubric and no rubric conditions on relative accuracy. However, the general rubric showed a trend towards improving absolute accuracy bias in both experiments, but not enough to be significantly different from the no rubric condition.

Why did the general rubric show no evidence for improving relative accuracy but showed a trend for improving absolute accuracy bias? Two possibilities may explain why
the general rubric could be improving students’ absolute accuracy bias. First, there might be a general response bias to the presence of the rubric. If a student believes that the inclusion of the rubric would be beneficial to their future test performance, they may be more likely to give themselves lower confidence judgments, and decrease their judgment magnitude in comparison to not having a rubric. In this case, having a rubric might make a student think about the difficulty that comes with understanding a text so that it might shift their judgments downward. In recent metacomprehension research exploring the use of diagrams, higher metacomprehension judgments were given for subjects given a diagram was during study compared to students not receiving diagrams, but this did not reflect overall performance (Serra & Dunlosky, 2010). Second, the general rubric may be allowing knowledge to become more accessible. By providing subjects with a second chance for study when making the judgments in addition to being provided a rubric to guide them to information in the text segments, this may be improving their accessibility to information. However, this did not change their relative accuracy, even though having the information from each text segment available from each text segment did not influence either conditions. In addition, Experiment 2 judgment magnitude only slightly decreased despite having knowledge more accessible in the rubric formats. Therefore, having knowledge more accessible within a rubric format may not account for the slight decrease in judgments in Experiment 2.

Overall, the rubric did not have an effect on either relative or absolute accuracy. Given that there was some promise of improving absolute accuracy bias with the general rubric, it seems that the rubric itself might need improvement to ensure that students are
using it effectively. Unlike the delayed-keyword generation task, the rubric task did not include any generative strategy (Thiede, 2003). The present version of the general rubric ties in information to relevant cues that the student should remember in order to receive full credit on future test questions, but it does not include any engaging aspect. Because the rubric from the present studies appears to be more passive than engaging, perhaps the inclusion of a pre-judgment recall might be beneficial to having an additional chance to retrieve information from the text, in addition to providing additional measures of metacomprehension accuracy (Nelson, Narens, & Dunlosky, 2004).

Because this was the first exploration of using rubrics to improve metacomprehension, one limitation to the rubric is that we only modeled it after the two formats specified by Allen and Tanner (2006). The general and specific rubrics were developed as a simplification of much more complex rubric examples that are used in classrooms. Future research with rubrics could incorporate versions that parallel what students might use in class for projects or assignments that contain several sub-categories of concepts that they must include in their final answer. Combining more complex rubrics with a generative task might be more effective for improving metacomprehension accuracy and could possibly benefit their overall performance as well.

In sum, because rubrics are used quite often in education, it is important to explore how they might be helping students’ metacomprehension accuracy for text materials. It is also important to investigate how different varieties of rubrics might contribute to improving metacomprehension accuracy. The current results demonstrate that rubrics in the form presented here are not effective at improving relative accuracy
any differently than not being given a rubric, but they might bring some promise to
improving their absolute accuracy bias to their overall performance. In future research, it
will be important to mirror rubrics used in the classroom in addition to incorporating a
generative task to explore how metacomprehension accuracy can be improved when a
rubric can be readily used by students.
References


<table>
<thead>
<tr>
<th>Variable</th>
<th>Question Type</th>
<th>Conditions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>No Rubric</td>
</tr>
<tr>
<td>Performance</td>
<td>Total</td>
<td>.40(.04)</td>
</tr>
<tr>
<td></td>
<td>Detail</td>
<td>.38(.04)</td>
</tr>
<tr>
<td></td>
<td>Concept</td>
<td>.42(.03)</td>
</tr>
<tr>
<td>Judgments</td>
<td>Total</td>
<td>75.33(.86)</td>
</tr>
</tbody>
</table>

Note: Data reported as means of the median test performance and median metacomprehension judgments with standard error of the mean in parentheses.
Table 2

*Test performance and metacomprehension judgments for Experiment 2*

<table>
<thead>
<tr>
<th>Variable</th>
<th>Question Type</th>
<th>Conditions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>No Rubric</td>
</tr>
<tr>
<td>Performance</td>
<td>Total</td>
<td>.41(.03)</td>
</tr>
<tr>
<td></td>
<td>Detail</td>
<td>.41(.04)</td>
</tr>
<tr>
<td></td>
<td>Concept</td>
<td>.41(.03)</td>
</tr>
<tr>
<td>Judgments</td>
<td>Total</td>
<td>70.57(2.83)</td>
</tr>
</tbody>
</table>

Note: Data reported as means of the median test performance and median metacomprehension judgments with standard error of the mean in parentheses.
FIGURE 1. Mean relative accuracy as a function of group and question type in Experiment 1. Error bars report standard error of the mean.
FIGURE 2. Mean absolute accuracy bias scores as a function of group in Experiment 1. Error bars report standard error of the mean.
FIGURE 3. Mean relative accuracy as a function of group and question type.

Error bars report standard error of the mean.
FIGURE 4. Mean absolute accuracy bias score as a function of group for Experiment 2. Error bars report standard error of the mean.
Appendix A

Sample questions from Thomas & McDaniel and modified questions used in the final test.

Kanchenjunga Mountain Text Questions

(Key: “D” = Detailed Questions; “C” = Conceptual Questions)

D. Mountains of Kanchenjunga are considered dangerous not only because of the ice avalanches but also because of the _________________.
   Answer: uncertain weather

C: Why is the Kanchenjunga mountain range more dangerous in comparison to other mountain ranges?
   Answer: Variable changes in the weather and ice avalanches make it far more dangerous.

D: The most striking characteristic of the ice in Kanchenjunga is its _________________.
   Answer: plasticity

C. In the Himalayas, snow is more quickly converted to glacial ice because:
   Answer: There is a large range in temperatures throughout the day.

D: The hanging glaciers in Kanchenjunga are best described as _________________.
   Answer: alive

C: Why is the Kanchenjunga ice more rubbery than the Alpine ice?
   Answer: Due to the way that the snow is changed almost immediately into ice.
D: The walls of ice in Kanchenjunga range from ______________ feet high.
   Answer: 600-1000

C: The downward speed of snow is much faster in Kanchenjunga than the Alps because:
   Answer: There is more snowfall and they are thicker.

D: What is ice like in the Alps?
   Answer: brittle

C: Avalanches are not as large in the Alps as on Kanchenjunga because:
   Answer: The ice is more brittle in the Alps and would break away more often and cause many smaller avalanches

D: On Kanchenjunga, ice walls _____ over the edge of cliffs.
   Answer: bend

C: How do avalanches commonly occur on Kanchenjunga?
   Answer: When ice bends over the edge of cliffs, it can crack when it can no longer handle the pull of gravity and other internal stresses.

Spider Text Questions

D: Spiders can range in size from fitting on the head of a pin to over ______________ inches in diameter.
   Answer: 14

C: Despite the great diversity of spider species, what similarities do all spiders share?
   Answer: They have the ability to make silk and have several sets of eyes.

D: Silken thread is so-fine spun yet strong that they have been used by ______________.
   Answer: scientists

C: Other than ballooning, list one benefit of spider’s silk for humans
   Answer: For science (optical instruments such as telescopes).
D: What time of year are spiders most likely to engage in ballooning?

Answer: Fall

C: How does a spider make a balloon?

Answer: It creates its own balloon with a long silk strand that can be carried along the wind.

D: How far from the closest shore have seamen spotted spider balloons?

Answer: hundreds of kilometers

C: What is the proof that spider’s balloons have been found beyond land?

Answer: Sailors at sea have spotted ballooning spiders hundreds of kilometers from the closest shoreline.

D: In Greek mythology, there was a young maid named ________.

Answer: Arachne.

C: Why did the Greeks create a myth about the spider?

Answer: They believed that spiders had magical powers.

D: Who changed Arachne into a spider?

Answer: Athena

C: Why was Arachne turned into a spider?

Answer: Arachne’s vanity that her spinning and weaving was the best made Athena upset, resulting in Athena turning her into a spider.
Appendix B

Rubric evaluation formats

General Rubric Format Example

<table>
<thead>
<tr>
<th>Full Credit</th>
<th>Half Credit</th>
<th>No Credit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contains all information related to the specific content and is well-defined. You have developed a strong understanding of the content area.</td>
<td>Contains some understanding related to the specific content but is lacking some critical components. You have developed an adequate understanding of the content area but may be missing some key components</td>
<td>Shows little to no understanding related to the specific content. There is little evidence that they have developed an understanding on the content area.</td>
</tr>
</tbody>
</table>

There are probably no other mountains more dangerous than those on Kanchenjunga for not only are there ice avalanches, but uncertain weather as well. The huge annual snow precipitation on Kanchenjunga plasters itself on the mountain and fills every hollow with ice. Because of this snowfall, plus the tug of gravity, these icy masses move downwards and feed the main glaciers.

How well will you perform on questions that test your understanding of the concepts in this text segment ( 0 = not likely to perform well at all to 100 = extremely likely to perform well)?

(Enter comprehension judgment # here)
Specific Rubric Format Example

Kanchenjunga Mountain Text

Specific Rubric Format

<table>
<thead>
<tr>
<th>Full Credit</th>
<th>Half Credit</th>
<th>No Credit</th>
</tr>
</thead>
<tbody>
<tr>
<td>You have a complete understanding that the Kanchenjunga mountains are more dangerous than other mountain ranges because of their variable weather changes and ice avalanches.</td>
<td>You have a partial understanding that the Kanchenjunga mountains are more dangerous than other mountain ranges but you may not understand completely why they are dangerous.</td>
<td>You have no understanding that the Kanchenjunga mountains are more dangerous than other mountain ranges.</td>
</tr>
</tbody>
</table>

There are probably no other mountains more dangerous than those on Kanchenjunga for not only are there ice avalanches, but uncertain weather as well. The huge annual snow precipitation on Kanchenjunga plasters itself on the mountain and fills every hollow with ice. Because of this snowfall, plus the tug of gravity, these icy masses move downwards and feed the main glaciers.

How well will you perform on questions that test your understanding of the concepts in this text segment (0 = not likely to perform well at all to 100 = extremely likely to perform well)?

(Enter comprehension judgment # here)