Question Format Biases College Students' Metacognitive Judgments for Exam Performance

Michael J. McGuire

Washburn University

Received: 04 August 2020; Accepted: 10 December 2021

College students in a lower-division psychology course made metacognitive judgments by predicting and postdicting performance for true-false, multiple-choice, and fill-in-the-blank question sets on each of three exams. This study investigated which question format would result in the most accurate metacognitive judgments. Extending Koriat's (1997) cue-utilization framework to these judgments, each format gave students different cues on which to base judgments. Further, each format has different probabilities of correctly guessing, which can skew accuracy. Students reported the lowest estimates for fill-in-the-blank questions. Accuracy measured using bias scores showed students' predictions and postdictions were most accurate for multiple-choice items. Accuracy measured using gamma correlations showed students' predictions were most accurate for multiple-choice items and postdictions were most accurate for fill-in-the-blank items. Based on the findings, educators are encouraged to consider what implications question format have on metacognitive processes when testing students over studied material. And, for researchers, the findings support the use of different accuracy measures to get a more detailed understanding of factors influencing metacognitive judgment accuracy.

By having students predict how well they will do on an exam and postdict how well they thought they just did, educators provide students with a chance to learn more about their metacognitive skill of monitoring performance. Without accurate monitoring, students may study ineffectively by either over- or underpreparing for upcoming exams (e.g., spending less time on poorly understood material and more time on well understood material). Likewise, the educator/researcher stands to learn in this exchange as well. Specifically, the educator can learn the extent to which question format, an oft-neglected variable in classroom studies, influences both estimated values (magnitude) and accuracy prior to and immediately following test taking.

When calculating metacognitive accuracy, measures usually fall into one of two broad categories – calibration or absolute, and relative or resolution (Dunlosky & Lipko, 2007; Maki, 1998; Nelson, 1996). Measures from both categories provide different lenses through which one can evaluate accuracy. Studies conducted in classroom contexts have generally reported absolute measures indicating better-than-chance accuracy for both predictions and postdictions with postdictions being more accurate (Maki & McGuire, 2002). Absolute measures indicate the degree of match between estimated and actual performance or the relationship between overall estimates and actual performance computed as a between-groups correlation. The closer a group's estimate is to actual performance, the better calibrated the estimate. The bias score, a common absolute measure, is calculated by subtracting one's estimated performance from her actual performance. The closer to zero the bias score, the more accurate the individual's judgment. Further, the score's sign indicates overconfidence (positive sign) or underconfidence (negative sign). A shortcoming of absolute measures involves discrimination of memorability levels among tested items within participants, which is corrected by using within-subjects measures that assess relative accuracy.

Relative accuracy, or resolution (Koriat, 1997; Maki, 1998; Nelson, 1996), indicates the relationship between one's judgments and performance on individual items or sets of items across groups of studied material (e.g., passages of texts). Unlike absolute measures, these measures indicate how well one can discriminate among items known well versus those known less well. For example, the closer to the value of one the Goodman-Kruskal gamma correlation, a common and preferred relative measure (Nelson, 1984, 1996) the more likely one's ratings correspond with actual outcome (i.e., items with higher ratings are more likely to be correct compared to items with lower ratings). Note this correlation says nothing about how well one's overall estimated performance compares to actual performance, which can be assessed using an absolute measure of accuracy such as the bias score. Even though one may demonstrate strong confidence as measured by bias scores, one may not be able to discriminate among items within a set, or vice versa. As is often the case, gamma is computed within individuals between one's ratings and performance for each item or sets of items.

Question format can be expected to influence metacognitive judgments for several reasons. First, extending Koriat's (1997) cue-utilization framework to metacognitive judgments in the current study, the process of reporting out one's estimated performance is inferential in nature and influenced by different cues. One cue type, intrinsic cues, include factors associated with judged material such as a priori beliefs concerning question difficulty. Related to this idea, Bonner (2013) reviewed various test-taking strategies students may attempt when completing various question formats noting how some question formats allow for different types of strategies compared to other question formats. A key distinction Bonner noted and investigated was one of multiple-choice (MC) formats versus constructed response (CR) formats such as fill-in-the-blank (FIB). For MC items students are afforded potential strategies unlikely to occur for CR items. For example, students may use a process of elimination to arrive at correct answers when completing MC items compared to either true/false (T/F) items or FIB items (assuming no word bank is provided). For both T/F and MC items, students may engage in test wise strategies inapplicable for FIB items. Further, FIB items unlike either T/F or MC require students to recall answers. As a result, one may estimate worse performance on FIB items compared to

https://doi.org/10.20429/ijsotl.2023.17115

either T/F or MC items given recall tasks are generally more difficult, or require more cognitive processing, than recognition tasks. In line with this reasoning, Thiede (1996) reported higher ratings (i.e., performance estimates) for recognition items compared to recall items. In another study, Maki, Willmon, and Pietan (2009) reported higher overall ratings for MC items compared to both essay and recall tests. In a classroom context, de Carvalho Filho (2009) reported higher ratings for MC items compared to short answer items. One's a priori knowledge concerning question difficulty contributes, in part, to this influence. But, properties inherent in the question in addition to one's a priori knowledge could contribute as well and have yet to be identified.

Notwithstanding content or question difficulty inherent in question stems, question format allows for different probabilities of correctly guessing an answer, which could skew accuracy for estimated performance (i.e., metacognitive judgments) as a result (Lichtenstein et al., 1982; McKenzie et al., 2001). The probability of correctly answering a T/F item due to chance is higher than it would be for a 4-choice alternative MC item. Consequently, the resulting discrepancy between estimated and actual performance on T/F items is arguably greater compared to MC items. As a result, the estimate's accuracy for the T/F items compared to MC items decreases. Lichtenstein et al. (1982) noted a similar outcome referred to as the Hard-Easy effect in which absolute accuracy for more difficult items resulted in overconfidence contrasted with underconfidence for easier items. Assuming FIB items are most difficult while T/F items are easiest in terms of correctly guessing (i.e., 50% chance of correctly guessing), T/F items should yield underconfident ratings whereas FIB items should yield overconfident ratings. For MC items, based on findings from both lab and naturalistic studies (Hacker et al., 2000; Hawker et al., 2016), students should be overconfident but not as overconfident compared to FIB items. In general, question formats associated with higher chances of being correct should result in higher estimates yet yield lower accuracy compared to formats associated with lower chances of being correct.

Rarely have naturalistic studies investigated both predictions and postdictions using more than one type of question format. The majority of studies investigating predictions or postdictions have been limited to MC items (Foster et al., 2017; Grabe et al., 1990; Hacker et al., 2000; Leal, 1987; Shaughnessy, 1979; Sinkavich, 1995). Moreover, a number of studies have investigated only predictions (Grabe et al. 1990; Leal, 1987) or only postdictions (Callender et al., 2016; de Carvalho Filho, 2009; Dutke et al., 2010; Hawker et al., 2016; Shaughnessy, 1979; Sinkavich, 1995). Hacker et al. (2000) is believed to be the only known study to investigate both predictions and postdictions in a classroom context though they only used MC items.

Of the naturalistic studies reviewed, only two used more than one question format. Callender et al. (2016) used MC items and "other" items defined as matching, T/F, and short answer items, yet the researchers did not investigate the effect different formats had on postdictions. In a classroom context, de Carvalho Filho (2009) investigated the effect of question format (MC items and short answer items) on metacognitive judgments in the form of postdictions but not predictions. The present study attempted to replicate and extend findings reported in de Carvalho Filho by investigating the extent to which question format (i.e., T/F, MC, FIB items) influences both predictive and postdictive ratings (a.k.a., magnitude) and accuracy using two different measures for accuracy.

The purpose of the current study was to answer the following three questions:

1. For which format do students provide the highest predictive and postdictive ratings (i.e., estimated performance in percentages)?

T/F items are expected to yield the highest ratings with FIB items being the lowest.

2. For which format are students most overconfident and underconfident?

T/F items are expected to yield underconfident ratings with the other formats associated with overconfidence. It is unclear which format will result in being most accurate (i.e., bias scores closest to zero).

3. For which format do students demonstrate highest relative accuracy (i.e., gamma correlations)?

Students should show higher gamma correlation values for FIB items compared to T/F items with MC items in-between the two. For each research question, differences between predictions and postdictions were also investigated.

METHOD

Participants

Thirty-nine students (28 females, 11 males) enrolled in the Psychology of Gender, a lower division course, at a medium-sized Midwestern university located in the U.S. consented to allow the author to analyze their data for this project.

Materials and Procedure

Prior to the beginning of the semester in which the study took place, I received IRB approval. During the first week of class, I informed all students about the study and informed them that the tasks associated with the current study were considered part of the course's requirements though they would not be graded on said activities. Next, I distributed consent forms. Students indicated on the forms whether they consented to allow me to analyze their responses on tasks for this investigation, folded the forms, and placed them in a manila envelope. I then reviewed the procedure for gathering predictions and postdictions noting that analysis of their predictions and postdictions would not occur until after grades were submitted and only for those students consenting for the study. Students were reminded that participation was completely voluntary and would not count towards their course performance in any way. After grades were submitted at the end of the semester, I opened the manila envelope to determine who consented to participate and thus whose data to analyze.

Predictions

On the day of each exam, immediately prior to the administration of the exam, students completed a form on which they predicted their exam performance for each type of question (i.e., T/F, MC, and FIB) on a scale ranging from 0% to 100%. The students' predictions represent the overall "magnitude" of their exam estimates listed in Table 1. They then folded their prediction forms in half and placed them in a manila envelope, which was then sealed for the duration of the semester. See Appendix A for the form used on the first exam. Forms used on subsequent exams were identical except for the chapter numbers listed.

Exams

After students returned prediction forms, they received the actual exam. See Appendix B for a sample template of the exam for one chapter of material. Students completed three exams each covering content from three chapters of the assigned textbook and corresponding lecture material. For each question format (i.e., T/F, MC, FIB), I wrote questions varying in difficulty (easy, medium, hard) and knowledge (factual, conceptual, and applied) levels. For each chapter covered on the exam, there were three sets of four questions per question format (i.e., four T/F items, four MC items, and four FIB items per chapter). Thus, there were a total of 36 items (3 chapters x 3 question formats x 4 question items). Here are sample questions taken from the first exam:

> T/F: Sex refers to nonbiological mechanisms when referring to the two different sexes.
>
> MC: Which of the following is a special sex tissue contained by both sexes of human fetuses during the first six weeks of uterine development, which has the potential to develop into either male or female sex structures?
>
> a. Amygdala
> b. Androgen
> c. Anlagen
> d. Testosterone
>
> FIB: ______ biases include our values, beliefs, attitudes, and assumptions.

Postdictions

After completing each section of an exam delineated by Chapter and Question Format, students made postdictions using the same scale used for predictions as noted in Appendix B. The students' postdictions represent the overall magnitude of their exam estimates listed in Table 1. To avoid potential conflict of interest, I announced to students that the results of their postdictions would not be analyzed until after grades were submitted. Thus, after I returned the completed exams for students to review I then placed exams in manila envelopes and sealed them until after grades were submitted at the end of the semester.

DESIGN AND DATA ANALYSIS

The overarching design of the current study formed a 3 (Question Format: T/F, MC, and FIB) x 2 (Judgment: predictions and postdictions) repeated-measures design. As such, three 3 x 2 repeated-measures ANOVAs were conducted to test the research hypotheses as there were three primary dependent measures of interest: students' estimated performance (a.k.a., overall magnitude of predictions and postdictions), students' level of accuracy in terms of over- versus underconfidence measured using bias scores, and students' level of accuracy in terms of resolution measured by calculating gamma correlations. Significant interactions were followed up with simple effects ANOVAs with alpha set to .025 to control for familywise error rates (all follow-up analyses employed Bonferroni corrections to control for familywise error rates in which alpha (.05) was divided by the number of analyses to be performed). Given there were three levels for Question Format (i.e., T/F, MC, and FIB), a significant simple effects analysis was followed up with pairwise comparisons setting alpha to .017.

Prior to starting each exam, students predicted their performance for each type of Question Format on a scale from 0% - 100% in terms of how confident they would perform. Students' overall magnitude was calculated by computing the average estimates (see means in Table 1) of performance for each question type across three exams for both predictions and postdictions. Their estimates provided a measure of perceived difficulty with higher estimates indicating better performance and thus easier items whereas lower estimates indicated worse performance and thus harder items.

Bias scores, or the difference in value between students' estimates and actual performance, provide two ways to evaluate metacognitive judgment accuracy. First, the sign of the bias score indicates whether students were either underconfident (negative sign) or overconfident (positive sign) with scores closer to zero indicating more accurate estimates. Thus, a score of zero indicates perfect accuracy (e.g., student predicts 85% and actually scores 85%). The averages of students' bias scores were computed across all three exams (see "Bias Score" row under the heading of 'Metacognitive Judgment Measure' in Table 1) as a function of Question Format and Judgment Type and submitted into a 3 x 2 RM ANOVA. Next, to determine the extent to which bias scores deviated from zero, and thus inaccuracy, I ran single-sample *t*-tests with zero as the test value. Significant *t*-tests indicate strong bias whereas nonsignificant *t*-tests indicate less bias. Strong bias suggests room for improvement in terms of accurately estimating performance.

In contrast to bias scores, which compare overall estimates to overall performance resulting in an absolute, or global measure of accuracy, gamma correlations (hereafter referred to gammas)

Table 1. Mean Values on Dependent Measures as a Function of Question Format and Metacognitive Judgment Type

	Question Format & Judgment Type					
	T/F Items		**MC Items**		**FIB Items**	
	Prediction	**Postdiction**	**Prediction**	**Postdiction**	**Prediction**	**Postdiction**
Metacognitive Judgment Measure	***M (SD)***	***M (SD)***	***M (SD)***	***M (SD)***	***M (SD)***	***M (SD)***
Magnitude	78.37 (10.23)	84.23 (9.64)	80.97 (9.17)	79.57 (12.24)	67.66 (14.00)	63.34 (17.65)
Bias Score	-12.27** (10.76)	-6.49** (9.71)	3.52 (13.32)	2.68 (13.00)	12.12** (20.15)	7.94** (20.15)
Gamma	.07 (.65)	.35** (.48)	.28* (.52)	.31** (.50)	0.04 (.45)	0.60** (.39)

* *p* < .01 (single-sample t-tests)
** *p* < .001

provide an accuracy measure of association between estimates and actual performance. Specifically, this correlational measure indicates the extent to which higher estimates (i.e., predictions and postdictions) correspond with better performance (i.e., actual performance). The resulting intracorrelations (i.e., correlation between estimated performance on sets of identically formatted items and performance on set of items) range from -1 to +1 with a value of positive one indicating perfect accuracy compared to correlations closer to zero, which suggest no relationship between estimates and actual performance. Thus, each student's estimates for individual sets of T/F items were correlated with their performance on those sets of T/F items and likewise for the other formats. A total of 18 data points (e.g., 9 T/F estimates and 9 T/F actual scores) were correlated for each student per question format. The gammas (see "Gamma" row under the heading of "Metacognitive Judgment Measure" in Table 1), calculated individually for each student, were submitted into a 3 x 2 RM ANOVA and followed up with simple effects ANOVAs for each judgment type as a function of question format. Whereas ANOVAs indicate mean differences, they do not directly measure accuracy. Thus, single-sample *t*-tests were conducted to see whether gammas significantly differed from zero. Significance in this case suggests better than chance accuracy. For all the single-sample *t*-tests, alpha was set to .008 to control for familywise error rates given that there were six means to test. All aforementioned analyses were conducted using SPSS v25.

RESULTS

Overview

The results are organized around the three research questions mentioned previously:

1. **Which question format do students perceive as the easiest?**
2. **For which question format are students more likely to be overconfident versus underconfident as indicated by bias scores? and**
3. **For which format are students most accurate as measured by bias scores and gamma correlations? Each research question was evaluated in terms of both predictions and postdictions.**

Which Question Format Do Students Perceive as the Easiest?

The averages of students' metacognitive judgment ratings in terms of overall magnitude for both predictions and postdictions (separately) were computed across all three exams as a function of Question Format (see row labeled "Magnitude" under the "Metacognitive Judgment Measure" heading in Table 1). Question Format significantly interacted with Judgment Type, *F*(1.62, 59.88) = 15.85, *MSE* = 38.05, *p* < .001, η_p^2= .30 (Greenhouse-Geisser corrected due to violation of sphericity). The significant interaction was followed up with two simple effects analyses to look at the effect of Question Format separately for each level Judgment Type. The simple effects analysis for students' predictions was significant, *F*(2, 36) = 42.46, *p* < .001, η_p^2= .702. An inspection of pairwise comparisons indicated that students rated MC items (*M* = 80.97, *SD* = 9.17) higher (and thus easier) than both T/F items (*M* = 78.37, *SD* = 10.23, *p* = .003) and FIB items (*M* = 67.67, *SD* = 17.65, *p* < .001). The simple effects analysis for students' postdictions was also significant, *F*(2, 36) = 51.67, *p* < .001, η_p^2= .740. In this case, however, students rated T/F items (*M* = 84.15, *SD* = 9.64) higher than both MC items (*M* = 80.13, *SD* = 12.24, *p* = .002) and FIB items (*M* = 63.48, *SD* = 17.65, *p* < .001).

For Which Question Format are Students Most Accurate as Assessed by Bias Scores?

Question Format significantly interacted with Judgment Type, *F*(1.62, 59.88) = 15.85, *MSE* = 38.05, *p* <.001, η_p^2= .30 (Greenhouse-Geisser corrected due to violation of sphericity). The significant interaction was followed up with two simple effects analyses to look at the effect of Question Format separately for each level Judgment Type on bias scores. The simple effects analysis for students' predictive bias scores was significant, *F*(2, 36) = 63.87, *p* < .001, η_p^2= .78. An inspection of pairwise comparisons indicated that students' bias scores for MC items (*M* = 3.52, *SD* = 13.32) were most accurate (closest to zero) and differed from both T/F items (*M* = -12.27, *SD* = 10.76, *p* < .001) and FIB (*M* = 12.12, *SD* = 20.15, *p* < .002). The simple effects analysis for students' postdictions was also significant, *F*(2, 36) = 51.67, *p* < .001, η_p^2= .642. The trend for postdictions was similar for predictions in which students' bias scores for MC items (*M* = 2.68, *SD* = 13.00) differed both T/F items (*M* = -6.49, *SD* = 9.71, *p* < .001) and FIB items (*M* = 7.94, *SD* = 20.15, *p* = .022). Thus, based on the signed differences, students' bias scores indicated students were underconfident for T/F items, slightly overconfident for MC items, and most overconfident for FIB items.

Students' predictive bias scores for T/F items significantly differed from zero, *t*(38) = -7.03, *p* < .001, *d* = 1.14 as did their postdictive bias scores *t*(38) = -4.31, *p* < .001, *d* = 1.12. Likewise, students' predictive bias scores for FIB items also significantly differed from zero, *t*(38) = 3.71, *p* < .001, *d* = .60 along with their postdictive bias scores, *t*(38) = 3.49, *p* = .001, *d* = .56. Students' predictive bias scores for MC items did not significantly differ from zero, *t*(38) = 1.63, *p* = .17, d = .27 nor did their postdictive bias scores, *t*(38) = 1.16, *p* = .26, *d* = .19. The result of bias scores for MC items not significantly differing from zero indicates students' accuracy was best for MC items compared to either T/F or FIB items.

For Which Question Format are Students Most Accurate as Assessed by Gammas?

Question Format significantly interacted with Judgment Type, *F*(2, 62) = 5.49, *MSE* = .18, *p* = .006, η_p^2= .15. The simple effects analysis for students' predictive gamma correlations was nonsignificant, *F*(2, 30) = 1.99, *p* = .154, η_p^2= .12 indicating that Question Format did not influence students' accuracy when making predictions. The simple effects analysis for students' postdictive gamma correlations was significant, *F*(2, 30) = 10.62, *p* < .001, η_p^2 = .42. An inspection of the pairwise comparisons indicated that students' postdictive gammas for FIB items (*M* = .60, *SD* = .39) were significantly higher than both T/F items (*M* = .35, *SD* = .48, *p* < .001) and MC items (*M* = .31, *SD* = .50, *p* = .005). Thus, students' gamma correlations indicated students were most accurate when postdicting performance FIB items compared to performance on both T/F and MC items.

Next, gammas were analyzed using single-sample *t*-tests to determine which gammas significantly differed from zero indicating better than chance accuracy. A summary of statistically significant findings is presented in Table 1 (see footnotes). Students predictions were better than chance for only MC items, $t(37) = 3.31, p = .002, d = .54$ whereas their postdictions were better than chance for all three question types: T/F items $t(31) = 4.14, p < .001, d = .73$; MC items $t(36) = 3.76, p < .001, d = .62$; FIB items $t(38) = 9.51, p < .001, d = 1.52$.

DISCUSSION

The current study investigated the effect of the format of exam questions on students' metacognitive judgments for exam performance in a classroom context. Students' predictions and postdictions were analyzed in terms of overall ratings indicating estimated percentages correct (i.e., magnitude of ratings). This measure provides information about students' perception of difficulty associated with question formats. In contrast to magnitude, accuracy was measured using two indices. Absolute accuracy was measured using bias scores to provide an index of how over- versus underconfident students were when predicting and postdicting their performance. And, relative accuracy was measured using gamma correlations providing an index of resolution or the degree to which students' ratings related to performance. This study is unique in that students' estimated performance was analyzed as a function of question format (i.e., T/F, MC, and FIB) and judgment type (i.e., predictions and postdictions) in a classroom context.

Based on overall magnitude of exam performance, students perceived FIB items as most difficult compared to T/F and MC items for both predictions and postdictions. This finding, in part, replicates one of de Carvalho Filho's (2009) findings in which students rated MC items more confidently than short-answer items. Several researchers have reported similar findings in the lab when comparing recognition with recall queries (Maki et al. 2009; Miesner & Maki 2007; Thiede 1996). Such findings are in accord with an extension of Koriat's (1997) Cue Utilization Theory. More specifically, metacognitive judgments are in part informed by one's knowledge concerning item difficulty (e.g., FIB items are harder than MC items) referred to as an intrinsic cue. Thus, individuals' a priori theory concerning question format (i.e., which format is easier) influences their metacognitive judgments, which is also in line with Flavell's (1979) concept of metacognitive knowledge (i.e., one's knowledge of factors influencing cognitive performance). Given the results along with previous findings concerning this distinction between question formats, educators can provide students with information in advance of exams in class. Students could then make a more informed decision as they set out to study for upcoming exams with different question formats. Perceived difficulty may impact testing anxiety with more difficult question formats increasing anxiety while testing. Knowing in advance the nature of the exam could potentially lower anxiety especially if practice, or sample, items were provided in advance of testing.

Students estimates of exam performance were also evaluated in terms of accuracy with the first measure of accuracy, bias scores, providing an index of over- versus underconfidence. Extreme over- and underconfident bias scores suggests there is room for improvement when estimating exam performance. The significant interaction between Question Type and Judgment indicated that students' predictions were more extreme especially when estimating performance for T/F items and overconfident for FIB items. Compared to postdictive bias scores, students performed better than expected on T/F items and worse than expected on FIB items. These findings are in accord with the Hard-Easy effect (Lichtenstein & Fischhoff, 1977; Lichtenstein et al., 1982). Accordingly, individuals are more likely to exhibit underconfidence for "easy" items and overconfidence for "hard" items. Thus, it is unclear whether students really did know more than they thought for T/F items, or whether the difference was due in part to correctly guessing, a likely factor. Students are most accurate at gauging how well they will do on MC items compared to T/F and FIB items prior to taking exams (predictive bias scores) and afterwards (postdictive bias scores). Thus, when using T/F and FIB items on exams, educators are urged to inform students about the potential biases that may affect their estimated scores which can in turn influence study methods. If students are planning to prepare for T/F items compared to FIB items, they may be better off allocating more effort towards preparing for FIB items relative to T/F items given the tendency to perform better on T/F items and worse on FIB items relative to their initial estimates. This recommendation assumes students know in advance which question formats will be on exams.

In addition to over- versus underconfidence, bias scores were tested against zero to see which question formats resulted in the best absolute accuracy. In this case, scores closer to zero indicate greater accuracy. Thus, if a score significantly differs from zero, then it would not be considered very accurate. Bias scores for T/F and FIB items significantly differed from zero, whereas scores for MC items did not significantly differ from zero. Thus, students were most accurate for MC items. Students presumably get more practice with MC items compared to the other formats. And, students' chances of correctly guessing are increased for T/F items relative to either MC or FIB items. Thus, there is an increased chance of getting items correct when one does not know the items. To ameliorate this chance of correctly guessing, educators are encouraged to provide students with formative assessments using T/F and FIB items for students to get a better sense of how well they will do. For example, in their lab study, Smith and Karpicke (2014) found retrieval practice to help students learn more effectively regardless of question format (i.e., MC, short-answer, and hybrid) compared to study alone. Studying in the format of retrieval practice may lead to less bias and more effective allocation of study time when preparing for these question formats.

An additional measure of accuracy, gammas, was calculated to assess students' relative accuracy (Nelson 1984, 1996) as a function of Question Format and Judgment Type. For this measure, postdictions significantly differed as a function of Question Format whereas predictions did not. Overall, students more accurately postdicted performance compared to predicted performance. The increased accuracy for postdictions replicated previous findings (Hacker et al. 2000; Maki et al. 2009; Pierce & Smith 2001). The better accuracy for postdictions may be taken as support for the Accessibility model (Koriat, 1993) given that more information on which to make estimates is provided for postdiction queries compared to prediction queries. Follow-up comparisons for postdictions indicated students' greater postdictive accuracy stemmed FIB items compared to both T/F and MC items.

Like bias scores, gammas can also be tested against zero. Unlike bias scores, statistical significance indicates greater than chance accuracy. In terms of predictions, students showed above chance accuracy for only MC items whereas for postdictions students showed above chance accuracy for all three question formats. This finding suggests that students are able to predict performance for MC items better than either T/F or FIB items in terms of relative accuracy. And, students have much better accuracy after taking exams. As noted earlier, if instructors are constructing exams using multiple question formats, attention should be given to formats other than MC items to give students a better sense of how likely they might perform. For example, using formative assessments utilizing question formats to be used on summative assessments is recommended. Doing so will provide students with practice so they can better gauge how well they will perform and thus contribute to their metacognitive skills for completing such items. Better yet, having students predict and postdict performance for low stakes practice quizzes/exams could inform students about their abilities and alert them to potential biases impacting their performance. Knowing more about what question formats will appear on upcoming exams can then help students study more effectively by practicing with those types of questions formats.

As with any naturalistic study, certain limitations are inherent, and some may be addressed in future studies. As is well known, random sampling (selection) increases external validity. Students in the present study were not randomly sampled. Thus, it is unclear whether the findings reported here will generalize to other college students. Another limitation involved a potential conflict of interest. I was the instructor of record and while students' predictions were secured until after grades were submitted, their postdictions were evident on returned exams. Thus, students may have responded differently than if the postdictions were kept confidential until after grades were submitted. To partially address this issue, students were informed that postdictions were considered part of the exam but were not graded. Additionally, students were informed that analysis of both predictions and postdictions would not be analyzed until after grades were submitted. To resolve this limitation in the future, researchers could run a similar study in a course for which they are not the instructor of record. The current sample consisted of psychology students, thus it's unclear how well these findings would transfer to different disciplines. Another limitation involved quality control of questions (e.g., item analyses). It is unclear how good, or poor, the questions were, which could have influenced both magnitude and accuracy of estimates. To address this issue de Carvalho Filho (2009) addressed this in his study by using questions from a test bank involving T/F and MC items. Previous research has also used stem-equivalent items (Bonner, 2013) for multiple question formats. In addition to controlling for difficulty, future research could also control for level of knowledge assessed (e.g., recognition vs. application), which is commonly reported in exam test banks. Another limitation was that the order of question format (i.e., T/F, MC, FIB) for predictions as well as postdictions remained constant. Would findings differ if the order was randomized for each exam? One could potentially randomize or counterbalance sections of question formats. Another limitation was not measuring students' academic aptitude (e.g., grade point average). Numerous studies have shown that stronger performing students generally show better metacognitive accuracy compared to weaker performing students (Hacker et al., 2000; Kruger & Dunning, 1999; Miller & Geraci, 2011). Though there are a number of limitations, this study provides future research with a basis on exploring the effect of multiple question formats on both predictive and postdictive metacognitive judgments and accuracy indices. Further, the results of this study underscore the importance of question format on metacognitive judgments, absolute accuracy of metacognitive judgments as measured by bias scores, and relative accuracy of metacognitive judgments as measured by gamma correlations.

CONCLUSION

This project endeavored to contribute to the lack of research addressing metacognition in a classroom setting involving metacognitive judgments and question formats. It is hoped that this investigation provides others with potential leads for further research as well as provide educators with ideas on how to address identified biases associated with question formats. Future research could look more closely at the properties of questions associated with either Bloom's Taxonomy (Anderson & Krathwohl, 2001) or Webb's Depth of Knowledge (Webb, 1997) in terms of what results in greater or lesser metacognitive awareness. Additionally, this investigation provides a different means (i.e., naturalistic manner) of assessing theory (e.g., Koriat's Cue-Utilization Model) used to understand metacognition. Ideally, an efficient theory of metacognition should accommodate findings from the lab and the classroom.

Going forward educators can use findings from the current study to inform their own instructional practices concerning test taking by informing students of potential metacognitive biases associated with different question formats, which is especially stronger for T/F and FIB items compared to MC items particularly when it comes to predicting future performance. Knowing that I, as a student, am likely to be less accurate when estimating future performance may in turn influence the manner in which I study for upcoming exams. In doing so, students should be able to more accurately gauge how well they think they have prepared for exams containing a variety of question formats thereby increasing their overall metacognitive awareness. I emphasize "should" in this previous sentence as this assertion is another avenue for future research. Now that we know certain question format types (especially T/F and FIB) bias students metacognitive judgments in terms of both magnitude and accuracy, what are the best ways to reduce the bias? One potentially helpful intervention for this question is a post-test analysis (Achacoso, 2004) or exam wrapper (Lovett, 2013) in which students not only estimate their performance but reflect on factors associated with their exam preparation and exam performance. Some research has shown that this type of intervention may provide students with the necessary insight to adapt their study strategies for improved exam performance (Edlund, 2020; Stephan et al., 2019; Trogden & Royal, 2019).

CONFLICT OF INTEREST

The author declares that he has no conflict of interest. All students completed informed consent forms. Only data from consenting students were analyzed for this project, which was IRB approved. Consent forms were kept confidential and reviewed the semester after the course ended to ensure grades were submitted prior to data analysis.

ACKNOWLEDGEMENTS

I thank my colleagues, Dr. Mike Russell, Dr. Cindy Turk, and Dr. Corey Zwikstra for reviewing my manuscript and for providing thoughtful feedback to help improve upon the organization and clarity of this paper.

CONTACT

Michael J. McGuire <michael.mcguire@washburn.edu>

REFERENCES

Achacoso, M.V. (2004). Post-test analysis: A tool for developing students' metacognitive awareness and self-regulation. *New Directions for Teaching and Learning*, *2004*(100), 115–119. https://doi.org/10.1002/tl.179

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

Bonner, S. M. (2013). Mathematics strategy use in solving test items in varied formats. *The Journal of Experimental Education*, *81*(3), 409–428. https://doi.org/10.1080/00220973.2012.727886

Callender, A. A., Franco-Watkins, A. M., & Roberts, A. S. (2016). Improving metacognition in the classroom through instruction, training, and feedback. *Metacognition & Learning*, *11*(2), 215–235. https://doi.org/10.1007/s11409-015-9142-6

de Carvalho Filho, M. K. (2009). Confidence judgments in real classroom settings: Monitoring performance in different types of tests. *International Journal of Psychology*, *44*(2), 93–108. https://doi.org/10.1080/00207590701436744

Dunlosky, J., & Lipko, A. R. (2007). Metacomprehension: A brief history and how to improve its accuracy. *Current Directions in Psychological Science*, *16*(4), 228–232.

Dutke, S., Barenberg, J., & Leopold, C. (2010). Learning from text: Knowing the test format enhanced metacognitive monitoring. *Metacognition and Learning*, *5*(2), 195–206. https://doi.org/10.1007/s11409-010-9057-1

Edlund, J. E. (2020). Exam wrappers in psychology. *Teaching of Psychology*, *47*(2), 156–161. https://doi.org/10.1177/0098628320901385

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American Psychologist*, *34*(10), 906–911. https://doi.org/10.1037/0003-066X.34.10.906

Foster, N. L., Was, C. A., Dunlosky, J., & Isaacson, R. M. (2017). Even after thirteen class exams, students are still overconfident: The role of memory for past exam performance in student predictions. *Metacognition and Learning*, *12*(1), 1–19. https://doi.org/10.1007/s11409-016-9158-6

Grabe, M., Bordages, W., & Petros, T. (1990). The impact of computer supported study on student awareness of examination preparation and on examination performance. *Journal of Computer-Based Instruction*, *17*(4), 113–119. Retrieved from psyh. (1991-19731-001)

Hacker, D. J., Bol, L., Horgan, D. D., & Rakow, E. A. (2000). Test prediction and performance in a classroom context. *Journal of Educational Psychology*, *92*(1), 160–170. https://doi.org/10.1037/0022-0663.92.1.160

Hawker, M. J., Dysleski, L., & Rickey, D. (2016). Investigating general chemistry students' metacognitive monitoring of their exam performance by measuring postdiction accuracies over time. *Journal of Chemical Education*, *93*(5), 832–840. https://doi.org/10.1021/acs.jchemed.5b00705

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, 77(6), 1121–1134. https://doi.org/10.1037/0022-3514.77.6.1121

Leal, L. (1987). Investigation of the relation between metamemory and university students' examination performance. *Journal of Educational Psychology*, *79*(1), 35–40. https://doi.org/10.1037/0022-0663.79.1.35

Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know? *Organizational Behavior and Human Performance*, *20*, 159–183.

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgments under uncertainty: Heuristics and biases* (pp. 306–334). New York: Cambridge University Press.

Koriat, A. (1993). How do we know that we know? The accessibility model of the feeling of knowing. *Psychological Review, 100*(4), 609–639. https://doi.org/10.1037/0033-295X.100.4.6

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General, 126*(4), 349–370. https://doi.org/10.1037/0096-3445.126.4.34909

Lovett, M. C. (2013). Make exams worth more than the grade: Using exam wrappers to promote metacognition. In M. Kaplan, N. Silver, D. Lavaque-Manty, & D. Meizlish (Eds.), *Using reflection and metacognition to improve student learning: Across the disciplines, across the academy* (pp. 18–52). Stylus.

Maki, R. H. (1998). Metacomprehension of text: Influence of absolute confidence level on bias and accuracy. In D. L. Medin (Ed.), *The Psychology of Learning and Motivation* (Vol. 38, pp. 223–248). https://doi.org/10.1016/S0079-7421(08)60188-7

Maki, R. H., & McGuire, M. J. (2002). Metacognition for text: Findings and implications for education. In T. J. Perfect & B. L. Schwartz (Eds.), *Applied Metacognition* (pp. 39–67). Cambridge University Press. https://doi.org/10.1017/CBO9780511489976.004

Maki, R. H., Willmon, C., & Pietan, A. (2009). Basis of metamemory judgments for text with multiple-choice, essay and recall tests. *Applied Cognitive Psychology*, *23*(2), 204–222. https://doi.org/10.1002/acp.1440

McKenzie, C. R. M., Wixted, J. T., Noelle, D. C., & Gyurjyan, G. (2001). Relation between confidence in yes–no and forced-choice tasks. *Journal of Experimental Psychology: General*, *130*(1), 140–155. https://doi.org/10.1037/0096-3445.130.1.140

Miesner, M. T., & Maki, R. H. (2007). The role of test anxiety in absolute and relative metacomprehension accuracy. *European Journal of Cognitive Psychology*, *19*(4–5), 650–670. https://doi.org/10.1080/09541440701326196

Miller, T. M., & Geraci, L. (2011). Unskilled but aware: Reinterpreting overconfidence in low-performing students. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(2), 502–506. https://doi.org/10.1037/a0021802

Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin*, *95*(1), 109–133. https://doi.org/10.1037/0033-2909.95.1.109

Nelson, T. O. (1996). Gamma is a measure of the accuracy of predicting performance on one item relative to another item, not of the absolute performance of an individual item: Comments on Schraw (1995). *Applied Cognitive Psychology*, *10*(3), 257–260.

Pierce, B. H., & Smith, S. M. (2001). The postdiction superiority effect in metacomprehension of text. *Memory & Cognition*, *29*(1), 62–67. https://doi.org/10.3758/BF03195741

Shaughnessy, J. J. (1979). Confidence-judgment accuracy as a predictor of test performance. *Journal of Research in Personality*, *13*(4), 505–514. https://doi.org/10.1016/0092-6566(79)90012-6

Sinkavich, F. J. (1995). Performance and metamemory: Do students know what they don't know? *Journal of Instructional Psychology*, *22*(1), 77.

Smith, M. A., & Karpicke, J. D. (2014). Retrieval practice with short-answer, multiple-choice, and hybrid tests. *Memory*, *22*(7), 784–802. https://doi.org/10.1080/09658211.2013.831454

Stephan, A., Whisler, L., Stephan, E., & Trogden, B. (2019). Using exam wrappers in a self-directed first-year learning strategies course. *2019 ASEE Annual Conference & Exposition Proceedings*, 33503. https://doi.org/10.18260/1-2--33503

Thiede, K. W. (1996). The relative importance of anticipated test format and anticipated test difficulty on performance. *Quarterly Journal of Experimental Psychology: Section A*, *49*(4), 901–918. https://doi.org/10.1080/027249896392351

Trogden, B. G., & Royal, J. E. (2019). Using exam wrappers in chemistry mathematics and statistics courses to encourage student metacognition. *Journal on Excellence in College Teaching, 30*(3), 71–96.

Webb, G. (1997). Deconstructing deep and surface: Towards a critique of phenomenography. *Higher Education*, *33*(2), 195–212.

APPENDIX A

Participant ID: __________

Instructions: First, fill in the blank for the Participant ID with any numeric, alphanumeric, or name you can think of. Remember this ID for your exam. There are nine slots for you to report your predictions. Each slot represents the question type and chapter for the question type. Please make a prediction on a scale of 0 – 100 indicating the percentage of questions you predict you will get correct. There are only four questions of a particular type (e.g., true-false) per chapter. Thus, if you write 0, 25, 50, 75, 100, then you are predicting that you will get 0, 1, 2, 3, or 4 correct, respectively. Please do not feel constrained by the scale. That is, use any number that corresponds to how much you feel you will get correct. For example, if you feel extremely confident that you will correctly answer 3 true-false questions for Chapter 1, but you might get 4, then you might put some number between 75 and 100 in the appropriate table cell. If you have any questions about this procedure, do not hesitate to ask me!!

{Respond from 0 – 100 (%)}	**Material**		
Question Format	**Chapter 1**	**Chapter 2**	**Chapter 3**
True-False			
Multiple-Choice (4-alternative)			
Fill in the Blank			

APPENDIX B

CHAPTER # MATERIAL

1.T F True-False Item 1

2.T F True-False Item 2

3.T F True-False Item 3

4.T F True-False Item 4

On a scale of 0 (absolutely nothing) to 100 (absolutely all), indicate how well you performed: _____

5. Multiple Choice Item 1
   a. Choice A b. Choice B c. Choice C d. Choice D
6. Multiple Choice Item 2
   a. Choice A b. Choice B c. Choice C d. Choice D
7. Multiple Choice Item 3
   a. Choice A b. Choice B c. Choice C d. Choice D
8. Multiple Choice Item 4
   a. Choice A b. Choice B c. Choice C d. Choice D

On a scale of 0 (absolutely nothing) to 100 (absolutely all), indicate how well you performed: _____

9. Fill-in-the-blank Item 1

10. Fill-in-the-blank Item 2

11. Fill-in-the-blank Item 3

12. Fill-in-the-blank Item 4

On a scale of 0 (absolutely nothing) to 100 (absolutely all), indicate how well you performed: _____