Differences Between Functional and Subjective Overconfidence in Postdiction Judgments of Test Performance

Matthew C. Shake[1] & Leah J. Shulley[1]

[1] Department of Psychological Sciences, Western Kentucky University, Bowling Green, Kentucky

United States of America

Correspondence: Dr. Matthew C. Shake, Department of Psychology Sciences, Western Kentucky University, 1906 College Heights Blvd. #21030, Bowling Green, KY, 42101, USA. E-mail: matthew.shake@wku.edu

Abstract

Introduction. Recent research has shown that students tend be overconfident when judging future performance on coursework, particularly students with lower academic ability. Some research suggests that these lower performing students are "doubly cursed" in that they are not only less capable of assessing their own performance, but also unaware of their own metacognitive deficits. In contrast, other research has suggested that while low performers are certainly less capable, they are quite aware of that deficit. The present study investigated this issue in the context of judgments made about *past* performance (i.e., *postdictions*) on tests.

Method. One hundred thirty participants from an Introductory Psychology university class completed postdiction judgments of performance and confidence after three exams. Analyses of variance were used to compare low versus high-performing students.

Results. Findings showed that low performing students were more likely to overestimate their past test performance, but were also less subjectively confident in the accuracy of those postdiction judgments. Additionally, while the tendency to overestimate past performance did not improve across multiple tests for the low performers, subjective confidence in those postdictions did, such that low performers became slightly more confident in their postdictions over time.

Discussion and conclusion. This research highlights the fact that low performing students are not good at assessing performance, even over repeated testing. While they seem to be aware of their poor metacognitive judgment, their confidence in those judgments may increase over time. These results and their implications for educators and for theories of metacognitive awareness are discussed.

Keywords: metacognition, overconfidence, test performance, performance assessment

Received: 02/20/14 *Initial acceptance: 03/18/14* *Final acceptance: 30/06/14*

Diferencias entre la sobre-confianza funcional y subjetiva en juicios posdicción de ejecución de pruebas

Resumen

Introducción. Investigaciones recientes han demostrado que los estudiantes suelen ser demasiado confiados al juzgar el desempeño futuro de los cursos, en especial a los estudiantes con la capacidad académica inferior. Algunas investigaciones sugieren que estos estudiantes de menor rendimiento están "doblemente malditos" ya que no sólo son menos capaces de evaluar su propio desempeño, sino también son menos conscientes de sus propios déficits metacognitivos. En contraste, otras investigaciones han sugerido que, si bien de bajo rendimiento son ciertamente menos capaces, son muy conscientes de ese déficit. El presente estudio investigó esta cuestión en el contexto de la apreciación de los resultados anteriores (es decir, juicios posteriores) en los exámenes.

Método. Ciento treinta participantes de una clase de introducción a la Psicología completaron juicios posteriores de rendimiento y confianza después de tres exámenes. Los análisis de varianza se utilizaron para comparar baja frente a los estudiantes de alto rendimiento.

Resultados. Los resultados mostraron que los estudiantes de bajo rendimiento eran más propensos a sobreestimar su desempeño en la prueba pasada, sino que también eran menos subjetivamente seguros de la exactitud de esos juicios posteriores. Además, mientras que la tendencia a sobreestimar el rendimiento pasado no mejoró a través de múltiples pruebas para los de bajo rendimiento, la confianza subjetiva en aquellos juicios posteriores hizo, de modo que bajo rendimiento se convirtieron en poco más de confianza en sus juicios posteriores en el tiempo.

Discusión y conclusiones. Esta investigación pone de relieve el hecho de que los estudiantes de bajo rendimiento no son buenos en la evaluación del rendimiento, incluso a través de pruebas repetidas. Aunque parecen ser conscientes de su falta de juicio metacognitivo, su confianza en dichas sentencias puede en aumento con el tiempo. Se discuten estos resultados y sus implicaciones para los educadores y para las teorías de la conciencia metacognitiva.

Palabras clave*:* Metacognición, exceso de confianza, rendimiento en las prueba, evaluación del desempeño.

Recibido: 20/02/14 *Aceptación inicial: 18/03/14* *Aceptación final: 30/06/14*

Introduction

Metacomprehension has been defined as the ability to think about and judge one's own learning or comprehension (Dunlosky & Lipko, 2007). Understanding the factors affecting metacomprehension accuracy has been of long-standing interest to educators and reseachers, in part because it has been implicated in successful learning and comprehension (Dunlosky, Hertzog, Kennedy, & Thiede, 2005; Thiede, Anderson, & Therriault, 2003). Some studies have focused on *relative accuracy*, which is the correlation between metacomprehension judgments of performance and actual performance. Others have focused on *absolute accuracy*, which is an assessment of the degree to which the judgments of performance are above or below actual performance. Unfortunately, in both cases, humans typically have poor accuracy when attempting to predict their own performance; in fact, we frequently overestimate our future performance in a variety of domains (Dunlosky & Lipko, 2007; Kruger & Dunning, 1999). Some of the most poignant examples of this miscalibration come from undergraduate college students, who generally greatly overestimate how well they will do on assessments such as tests. This tendency to predict that one will do better than one actually does has been termed *functional overconfidence* (Hacker, Bol, Horgan, & Rakow, 2000; Maki, Shields, Wheeler, & Zacchilli, 2005; Miller & Geraci, 2011; Nietfeld, Cao, & Osborne, 2005). This problem leads to significant challenges for educators who, when checking for student comprehension, may make the errorneous assumption that students are able to effectively make judgments about their own abilities.

While a variety of individual differences appear to influence the accuracy of metacomprehension judgments (e.g., verbal ability, SAT scores, text difficulty; Kelemen, Winningham, & Weaver, 2007; Maki et al., 2005), one of the strongest predictors of functional overconfidence is academic performance. Specifically, students with the lowest class grades tend to have greater functional overconfidence as compared to those with the highest class grades (Hacker et al., 2000; Krueger & Mueller, 2002; Miller & Geraci, 2011). In other words, academically poorer students appear to be more likely to overestimate when predicting their upcoming performance. The reasons for this deficit remain unclear. One view argued by some researchers is that the cause is a lack of awareness of one's own poor metacomprehension; that is, low performers are "double cursed" because they are (a) less capable of assessing their own performance, but also (b) unaware of their poor metacognitive skills (Ehrlinger, Johnson, Banner, Dunning, & Kruger, 2008; see also Dunning, Johnson, Ehrlinger, &

 Electronic Journal of Research in Educational Psychology, *12*(2), 263-282. ISSN: 1696-2095. 2014, no. 33
http://dx.doi.org/10.14204/ejrep.33.14005

Kruger, 2003). Another view is that the high performers are simply closer to the ceiling and thus have less room to vary their judgments compared to the low performers (Krueger & Mueller, 2002). Most recently, Miller and Geraci (2011) argued that low performing students simply make guesses they deem reasonable, which turn out to be grossly inaccurate (Miller & Geraci, 2011). Arguing against the double curse account, Miller and Geraci (2011) provided evidence that functional overconfidence in low performing students is most likely not because they are unaware of their own metacomprehension deficits. In their findings, while low performing students were certainly more functionally overconfident in their prediction judgments, they were also less *subjectively confident* than high performing students in the accuracy of those prediction judgments of test performance, which suggests that these students are somewhat aware of their own poor metacognitive calibration (also see Dunlosky, Serra, Matvey, & Rawson, 2005). The present study had two specific aims: first, to assess which of the aforementioned views applies when students are making after-the-fact, postdiction judgments; second, to assess whether such functional and subjective confidence changes over time across several exams in a semester-long class. We briefly review literature in these areas next.

Postdiction Accuracy and Confidence in Postdiction Judgments

While the primary focus of the previously mentioned research was on the functional and subjective confidence in predictions of future performance, students also often make retrospective, after-the-fact judgments of performance. These *postdictions* can serve several useful purposes, such as helping to guide the student for future study efforts. While there is a substantial body of literature on prediction accuracy, there is comparatively less on the accuracy of metacomprehension postdictions (Hacker, Bol, & Keener, 2008). The reason for this is likely because prediction research is often seen as a way to potentially proactively alter future study behavior. For example, it is often argued that if prediction accuracy can be improved, it could theoretically alter future self-regulated learning in a way that would maximize study efforts and comprehension outcomes. However, postdiction accuracy also has an important relationship to study efforts and outcomes. Postdiction judgments allow a learner to retroactively judge their actual performance, which is a self-evaluative form of feedback that can alter perceptions and future effort (Glenberg & Epstein, 1985; Hacker et al., 2000; Maki & McGuire, 2002; Pierce & Smith, 2001). Dunlosky, Rawson, and Middleton (2005), for example, recently found that attempting to retrieve a test item *prior* to making a metacognitive

judgment (i.e., a form of postdiction judgment) boosted the subsequent accuracy of performance judgments.

While it is clearly important to understand the accuracy of postdictions, it is also important to understand how confident students are in their own postdictions. If a poor performing student is both overestimating their past performance *and* is extremely confident in the accuracy of that metacognitive judgment, it would suggest that these students are not only unskilled, but also unaware of their inability to metacognitively assess their academic performance. Recent research on this issue has focused primarily on predictions rather than postdictions, and the findings have been mixed. One account has argued that academically poorer students are "double cursed" (Ehrlinger et al., 2008) in that they are unskilled and unaware; however, other researchers have found that poor performing students are unskilled but very much aware of their poor metacognitive judgments (Miller & Geraci, 2011). One of the aims of the present study was to investigate this issue by providing evidence to help adjudicate between these two competing explanations, with a particular focus on postdiction accuracy and confidence judgments.

Postdiction Accuracy and Confidence Judgments Over Time in the Classroom

Most educators and students understand that metacognitive judgments are not static; rather, they can evolve over time with additional experience or practice. Indeed, postdiction judgments about performance are "experience-based" (Koriat, 1997, p. 367) and can alter how future effort is allocated (Stine-Morrow, Gagne, Morrow, & DeWall, 2004), as well as how accurate subsequent prediction judgments are. For example, it appears that the accuracy of prediction judgments improves over time in laboratory studies, presumably because those judgments become increasingly influenced by past judgments and test experience (Hertzog, Dixon, & Hultsch, 1990; Koriat, 1997). Unfortunately, most research in this area has used laboratory-based tasks; comparatively little work has examined postdiction judgments over time in real world classroom settings (Maki & McGuire, 2002). One study (Hacker et al., 2000) that did examine postdiction judgments over time found that academic performance was related to metacomprehension judgment accuracy over time in a semester-long college class. In that study, Hacker et al. (2000) had students make both prediction and postdiction judgments about test performance multiple times throughout a semester-long class. The researchers predicted that (a) postdictions should be more accurate than predictions (this hy-

pothesis was confirmed), and that (b) postdictions should improve over time. However, their data showed that low performing students did not improve in either predictions or postdictions over time, whereas high performers improved in both, especially postdiction accuracy (also see Bol, Hacker, O'Shea, & Allen, 2005).

However, the aforementioned studies did not examine subjective confidence in students' postdictions. In their recent research, Miller and Geraci (2011) examined primarily prediction accuracy and subjective confidence judgments (CJs) in those predictions; however, in their second experiment they did ask for both a prediction *and* postdiction CJ (on the last exam only). On the last exam, they found that low performers were still less confident in their prediction accuracy (i.e., subjective *under*confidence), though the difference in confidence from prediction to postdiction did not vary by academic performance (i.e., the decline from pre- to postdiction subjective confidence was similar in both high and low performers). None of the aforementioned studies have examined change in subjective confidence over time, and it remains unclear whether low performers are more or less likely to experience changes over time in their subjective confidence regarding postdiction judgments.

Present Study

To more fully elucidate these questions of objective (i.e., functional) postdiction accuracy and subjective confidence in those metacognitive postdictions, we asked students in multiple Introductory Psychology classes to make postdictions regarding their exam performance and to rate their confidence in those postdictions. They did so for three exams over the course of an entire semester. Given earlier research, we hypothesized that low performing students would show poorer postdiction absolute accuracy (i.e., functional overconfidence) as compared to high performing students. With regards to students' subjective confidence in their postdictions, if the data showed high confidence judgments in the low performers, it would bolster the "double cursed" account (Ehrlinger et al., 2008); on the other hand, if low performing students showed lower confidence judgments, the data would provide support for the "unskilled but aware" account (Miller & Geraci, 2011).

With regard to changes in these factors over multiple exams, if the data are consistent with Hacker et al. (2000), low performers should show less improvement in the absolute accuracy of their postdictions over time, as compared to high performers. It is less clear what to

predict regarding subjective confidence in the postdictions over time; however, if low performing students showed continued low subjective confidence in the postdiction judgments across multiple exams, it would lend credence to the hypothesis that these students are unskilled at making metacognitive judgments of performance, yet quite aware of that metacognitive deficit.

Method

Participants

One hundred thirty (130) students from five sections of an Introductory Psychology college class at a small university in the United States participated in the study. Sixty-five percent of the students were female. All sections were highly similar, insofar as they were taught by the same instructor, had the same instruction, the same course requirements, the same topics covered in the same sequence, and the same grading scheme.

Materials and Procedure

Over the course of each semester-long class, three non-cumulative multiple-choice exams were given. The exams were spread out evenly over the length of the semester, and each exam had 65 multiple-choice questions. Each exam covered five topics, and the questions on each exam were taken from a well-normed test bank. At the end of each exam, participants were asked to write down answers to two written questions. The first question asked: "How many of the questions on this exam do you think you answered correctly? Mark that number on the line below. For example, if you think you answered all 65 of the questions correctly, you would enter 65 on the line below." The second question asked: "How confident are you that you have correctly predicted your score on this exam?[1] Mark the number that corresponds to your confidence level below." Participants were shown a Likert-style scale from one to nine, with a one labeled as "Not at all Confident" and a nine labeled as "Very Confident." Students did not receive an incentive for completing these questions.

Data analysis

Twenty-one subjects (16%) were removed prior to data analysis for missing data (i.e., they chose not to answer both the postdiction and confidence questions).[2] For ease of comparison to other studies (e.g., Hacker et al. 2000; Miller & Geraci, 2011), and in order to more directly examine the lower and higher performing students in the sample, the data from the remaining

participants (N=109) were divided into Quartiles on the basis of actual exam performance. Postdiction absolute accuracy was calculated as the absolute difference between actual performance and postdicted performance. Means for actual number correct, postdiction judgment, absolute difference, and subjective confidence judgments (CJs) in the postdictions are all shown in Table 1. Note that positive numbers in the absolute difference scores reflect functional overconfidence, whereas numbers closer to zero reflect good metacognitive calibration (i.e., good ability to judge performance after the test). Omnibus *F* tests were used to compare between-group differences in postdiction absolute accuracy and confidence judgments. Mixed analyses of variance (ANOVA) were used to analyze any potential interactions of within-subjects differences (Exam) with between-subjects differences (Quartile).

Results

Absolute Accuracy (Functional Confidence) for Postdiction Judgments

Exam 1. An omnibus *F* test indicated a significant difference in postdiction absolute accuracy by Quartile, $F(3, 105) = 5.73$, $p < .01$. Tukey post hoc tests showed that the bottom quartile (Quartile 1) was significantly less accurate in estimating their test score than the students in the top two quartiles (both $p < .05$; also marginally less accurate than the next highest Quartile 2, $p = .09$). In other words, low performing students differentially overestimated how well they would do on the exam. Differences between the top three quartiles (Quartiles 4, 3, and 2) were *ns*.

Exam 2. Results for the second exam again showed a significant difference in postdiction absolute accuracy by Quartile, $F(3, 105) = 12.91$, $p < .001$. Tukey post hoc tests revealed a similar pattern to that of Exam 1: students who performed worse on the exam were more overconfident in their postdiction judgments (reflected in significant differences compared to Quartiles 4 and 3, ps $< .05$, and marginal compared to Quartile 2, $p = .08$). There were some reliable differences between the top 3 quartiles (see Table 1 for Means): students in the top quartile (Quartile 4) were significantly more accurate than only the bottom two quartiles (both ps $< .01$); in addition, Quartile 3 was not significantly more accurate than the quartiles above and below it (i.e., Quartiles 2 and 4, both ps $> .05$).

Exam 3. Results for the third exam replicated the findings from the first two exams, reflected in a significant difference in postdiction absolute accuracy by Quartile, $F(3, 105) = 13.58$, $p < .001$. Here, students in the bottom quartile were again more functionally overconfident than students in the three higher quartiles, as shown by positive absolute difference scores (all *p*s < .05). Comparisons between the three higher quartiles revealed a pattern similar to Exam 2 (i.e., Quartile 4 > 2 and 1, both *p*s < .01, Quartile 3 = 2 and 4, *ns*).

Table 1. **Means (and Standard Errors) for Number Correct, Postdiction, Absolute Accuracy, and Confidence Judgments by Exam and Quartile**

Quartile	Number Correct	Postdiction	Absolute Accuracy	Confidence
		Exam 1		
1	32.00 (0.93)	38.96 (1.64)	6.96 (1.99)	5.28 (0.38)
2	38.87 (0.87)	44.63 (1.18)	5.77 (1.49)	4.83 (0.28)
3	43.36 (0.84)	47.04 (1.45)	3.68 (1.34)	5.68 (0.23)
4	53.00 (0.77)	54.77 (1.04)	1.77 (1.15)	6.62 (0.22)
		Exam 2		
1	35.00 (1.23)	42.04 (1.71)	7.04 (1.75)	5.08 (0.35)
2	41.50 (0.78)	45.53 (1.09)	4.03 (1.30)	5.70 (0.26)
3	48.79 (0.73)	49.54 (0.95)	0.75 (1.05)	6.29 (0.25)
4	54.12 (0.62)	53.54 (0.86)	-0.58 (0.69)	6.38 (0.28)
		Exam 3		
1	34.48 (1.19)	41.56 (1.10)	7.08 (1.29)	5.76 (0.31)
2	44.77 (0.70)	45.47 (0.98)	0.70 (1.07)	5.47 (0.32)
3	48.29 (0.98)	47.75 (1.11)	-0.54 (1.36)	5.54 (0.33)
4	53.73 (1.00)	53.27 (1.06)	-0.46 (1.00)	5.96 (0.34)

Note. Maximum possible number correct on each exam was 65. Absolute Accuracy refers to the average raw difference between Number Correct and Postdiction; positive numbers indicate overestimates. Quartiles are based on average performance across all three exams.

Subjective Confidence in Postdiction Judgments

*Exam 1***.** An omnibus *F* test showed that students' subjective confidence (as measured by the Likert-scale rating) in their postdiction varied by quartile, $F(3, 105) = 9.14$, $p < .001$. The poorest performing students (both Quartiles 1 and 2) were significantly less confident in their metacognitive judgments than the better performing students (both Quartiles 3 and 4), all *p*s < .05, indicating that they were not as confident in the accuracy of their postdictions regarding exam performance. The difference between Quartiles 1 and 2, and between Quartiles 3 and 4, were *ns*.

Exam 2. Results for the second exam showed a pattern similar to Exam 1, with students' confidence varying by quartile, $F(3, 105) = 4.13$, $p < .01$. The students in the lowest quartile remained significantly less confident than those in the highest quartiles, $p < .01$. All other comparisons were *ns*.

Exam 3. In contrast to the first two exams, for the last exam of the class, there was no difference in confidence judgments as a function of Quartile exam performance, $F(3, 105) < 1$, *ns*.

Changes in Postdiction Absolute Accuracy and Confidence Judgments Over Time

The aforementioned results broken down by each exam showed that the accuracy of participants' postdictions seemed to vary by actual exam performance on all three exams, whereas their confidence in those postdictions only varied for the first two exams and not the last. It therefore seemed plausible that these metacognitive components could have varied over the course of the semester. That is, we next analyzed whether (a) the absolute accuracy of postdiction judgments change over repeated testing, and whether (b) confidence in those judgments change over repeated testing. While the between-exam data (see Table 1) seem to suggest that the bottom quartile students do not become more functionally accurate in their postdictions, this does not address within-person change since students are not necessarily in the same quartile for each exam. Indeed, when we examined whether there was relative rank-order stability in exam performance, spearman rank-order correlations showed that while students tended to score in similar quartiles across the three exams (correlations between quartile ranks ranged from .52 to .66, all *p*s < .001), the correlations were nowhere near perfect (i.e., +1.0). We therefore calculated an average across all three exam scores for each student and

then quartiled the sample based on this overall average, as other researchers have done (see Hacker et al., 2000).[3]

Absolute accuracy. Two 3 (Exam) x 4 (Quartile) mixed ANOVAs were run with absolute accuracy for the three exams as a within-subjects factor and Quartile as a between-subjects factor: one for postdiction absolute accuracy and one for confidence judgments. For postdiction absolute accuracy, the main effect of Quartile was significant, $F(3, 105) = 9.09$, $p < .001$, $\eta_p^2 = .21$, reinforcing the earlier result that lower quartiles tended to be more overconfident (Quartile 1: $M = 7.03$, $SE = 1.01$; Quartile 2: $M = 3.50$, $SE = .92$, Quartile 3: $M = 1.30$, $SE = .95$, Quartile 4: $M = .24$, $SE = .99$). The main effect of Exam was also significant, $F(2, 210) = 6.68$, $p < .01$, $\eta_p^2 = .06$, and the pattern of results indicated that the absolute accuracy of the postdiction judgments did improve across exams (Exam 1: $M = 4.54$, $SE = .76$; Exam 2: $M = 2.81$, $SE = .63$, Exam 3: $M = 1.70$, $SE = .60$), with pairwise comparisons showing that Exams 3 and 2 were better than the Exam 1 (ps < .01 and .05, respectively). The comparison of Exam 3 to Exam 2 was in the predicted direction, but not statistically reliable, $p = .12$. While these main effects did not interact, $F(6, 210) = 1.30$, *ns*, Figure 1 clearly indicates that the academically poorest students (Quartile 1) showed no improvement in absolute accuracy across the three exams (all pairwise t-test comparisons were *ns*). In contrast, the academically better students, particularly in the top quartiles, were approaching near perfect calibration by the second exam (pairwise comparison of Exam 1 to Exam 2 for the top quartile only showed $p = .05$).

Electronic Journal of Research in Educational Psychology, *12*(2), 263-282. ISSN: 1696-2095. 2014, no. 33
http://dx.doi.org/10.14204/ejrep.33.14005

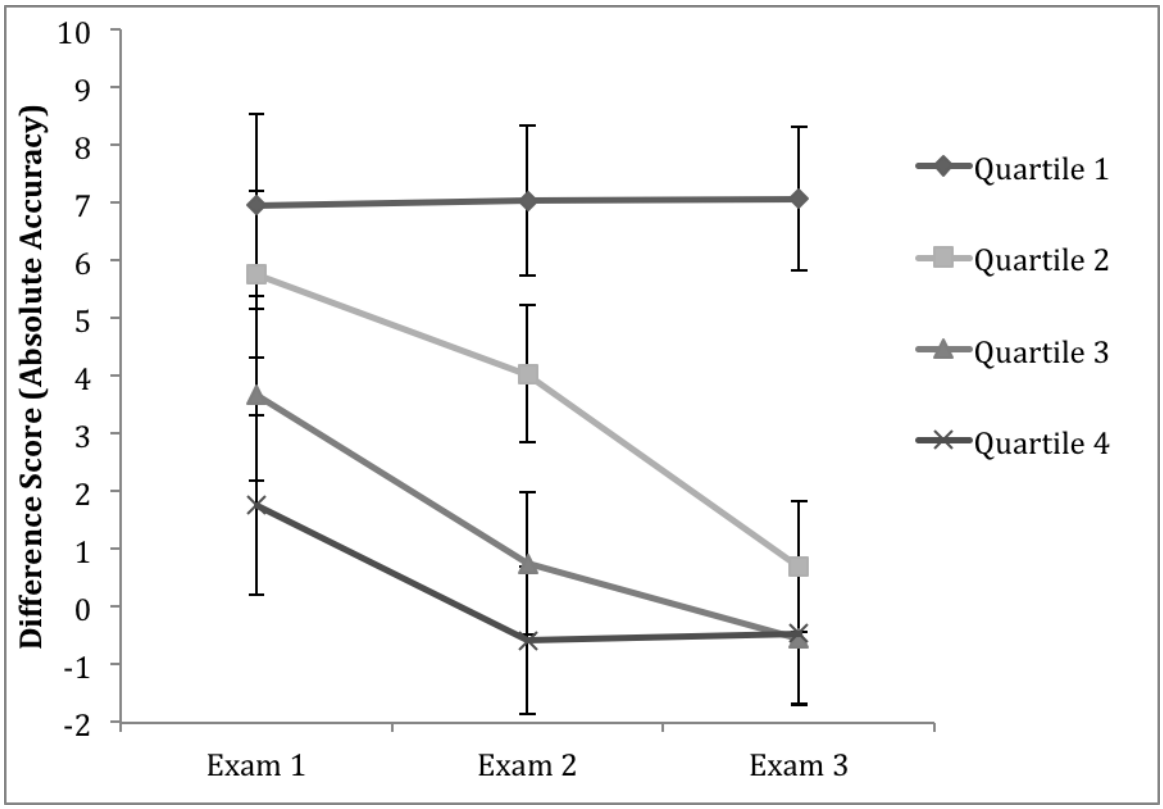

Figure 1. **Average Difference Scores (Absolute Accuracy) by Exam for each Quartile. Higher numbers indicate poorer accuracy in postdiction judgments.**

Confidence judgments (CJs). The ANOVA on confidence ratings indicated that collapsed across quartiles (i.e., all participants), there was no difference in confidence ratings across the three Exams, $F(2, 210) = 1.55$, *ns* (Exam 1: $M = 5.60$, $SE = .14$; Exam 2: $M = 5.86$, $SE = .14$, Exam 3: $M = 5.68$, $SE = .16$). However, the main effect of Quartile was significant, $F(3, 105) = 3.61$, $p < .05$, $\eta_p^2 = .09$ (Quartile 1: $M = 5.37$, $SE = .25$; Quartile 2: $M = 5.33$, $SE = .23$; Quartile 3: $M = 5.83$, $SE = .24$; Quartile 4: $M = 6.32$, $SE = .25$), again showing that academically poorer students were less confident in the accuracy of their postdiction judgments (Quartile 4 greater than Quartiles 1 and 2, both *p*s < .01; all other comparisons *ns*).

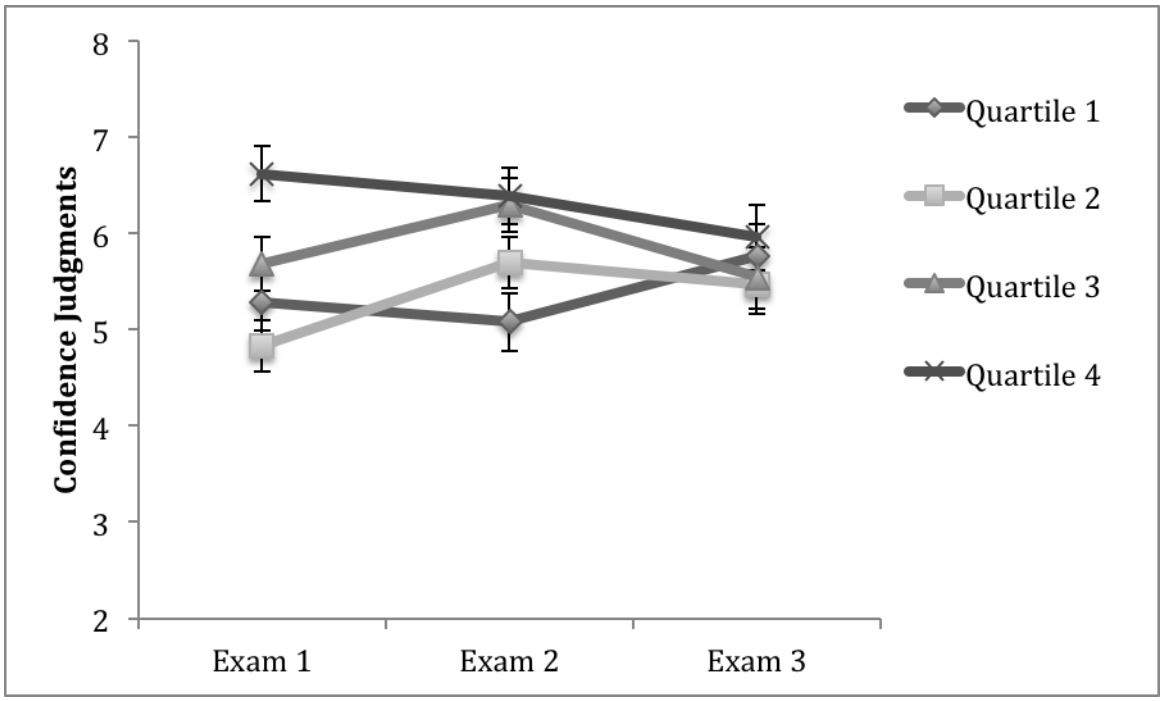

Figure 2. **Confidence Judgments by Exam for each Quartile. Higher numbers indicate greater confidence in the accuracy of the postdiction judgment.**

More importantly, confidence ratings showed an interaction of Exam and Quartile, $F(6, 210) = 3.70$, $p < .01$, $\eta_p^2 = .10$. This interaction, shown in Figure 2, shows that the bottom Quartile students were significantly less confident than the top Quartile students only for the first two Exams ($t(49) = 3.06$, $p < .01$ and $t(49) = 2.93$, $p < .01$, for Exams 1 and 2 respectively), but confidence was similar for the last Exam, $t(53) < 1$, *ns*. Within-subject comparisons indicated that top Quartile students tended to have stable confidence over time (all *ps* > .05, *ns*), whereas bottom Quartile students increased their confidence from Exam 2 to Exam 3 ($t(24) = 2.06$, p = .05; other comparisons were *ns*).

Discussion and Conclusions

Overestimating one's performance on a test is a significant and pervasive problem, particularly for educational settings. The present study showed that low performing students were significantly more overconfident than high performers with regard to the absolute accuracy of their postdiction (i.e., after testing) judgments of performance. This low performing overconfidence was present on each of three exams across an entire semester. This finding extends

Electronic Journal of Research in Educational Psychology, *12*(2), 263-282. ISSN: 1696-2095. 2014, no. 33
http://dx.doi.org/10.14204/ejrep.33.14005

other research on prediction (before testing) metacomprehension (e.g., Miller & Geraci, 2011) and indicates that the act of retrieving the test information does not significantly alter the metacognitive miscalibration that low performing students demonstrate. Taken together, this body of research suggests that academically poor students suffer from consistently poorer metacomprehension ability both with regards to their ability to predict upcoming performance and to retroactively evaluate past performance.

Absolute Accuracy of Postdictions

The present data also suggest that repeated testing within a classroom context may lead to improved postdiction accuracy, but only among higher performing students. In the present study, it was clear from Figure 1 that low academic performers did not improve their postdiction accuracy across exams. In that regard our data are consistent with Hacker et al.'s (2000) findings, who used a median split and also found no improvement among low performers in postdiction accuracy; we note that our data used a quartile split and thus may be a more sensitive representation of truly low performing students (i.e., comparison of the top and bottom quartiles as opposed to a 50/50 median split). The fact that the majority of students (i.e., those in the higher quartiles) tended to improve their absolute accuracy over the course of the semester, but that within each exam the low performing subjects still remained functionally overconfident, highlights just how strongly poor metacomprehension accuracy is intertwined with actual academic performance.

There are several potential explanations as to why the absolute accuracy of postdiction judgments might change across exams. One possibility is that it is a measurement artifact (Krueger & Mueller, 2002), such that high-performing students are already close to perfect (i.e., difference scores for absolute accuracy close to zero) and low performing students have more room to improve over time. Our data do not seem entirely consistent with this explanation, however. For example, while it is certainly the case that on all three exams, the top quartile was well-calibrated (mean difference scores ranged from 1.77 to -0.46), the top quartile also was never close to ceiling in actual performance (mean number correct ranged from 53-54 out of 65 possible). Additionally, as we noted earlier, the poorest performing students (Quartile 1) did not improve in their postdiction accuracy over time. It is also plausible that the global nature of the postdictions, in contrast to item-specific postdictions, could be a factor, since item-specific metacognitive judgments do tend to be more accurate (Dunlosky & Lipko, 2007). Other possible explanations for changes in postdiction accuracy include differ-

ences in the content of the tests, or individual patterns of student engagement with the test material (i.e., the better performing students were better at engaging with the material to be learned). Our data cannot adjudicate these accounts; future research should examine these possibilities with an experimental analog that more tightly controls materials and controls for some individual differences.

Subjective Confidence in Postdictions

An important issue for researchers and educators is whether learners are aware of their own metacomprehension deficits, and whether such awareness may vary over time or assessments. The present study found that reliable differences in confidence judgments regarding postdiction accuracy varied as a function of actual exam performance for the first two exams, but not for the last exam. On the first two exams, academically poor students were less confident in the accuracy of their postdictions than the high performing students, suggesting that these students may be at least somewhat aware of their poor metacomprehension abilities. The lack of effect for the last exam in the present data appears at first to conflict with Miller and Geraci (2011), who found that low performing students remained less confident than high performers on a second exam later in the semester. It should be noted, however, that Miller and Geraci asked students to make their post-test confidence judgment about a *prediction* judgment made before the exam (i.e., to revise a prediction guess after actual retrieval). It is plausible that low performers in this case, reflecting on their pre-test predictions, were less confident than they would be in the present study, where the CJ was made only regarding a postdictive judgment. In the present study, subjects were not forced to make and confront pre-test predictions, which could have unforeseen effects on subsequent *a posteriori* postdictions.

Confidence judgments may also change over the course of multiple exams, as reflected in an interaction between exam and semester-long academic performance. In particular, the present data indicated that low performers may actually become somewhat more confident in the accuracy of their postdictions over time, that is, increased confidence on the third and final exam (Figure 2). The reason for this change is not entirely clear but several possibilities exist. For example, topic material may have played a role. In all sections of the Introductory Psychology course, topics covered for the final exam were topics that tend to be deemed more popular and intrinsically interesting to the typical undergraduate: for example, abnormal psychology, personality, and social psychology. Perhaps the low-performing students felt more confident in their postdictions as a function of the topics covered; future re-

search could address this with a design that counterbalances the order of topics in the course. Another possibility is that the change in confidence for low performing students could be due to slightly greater study time for the last exam (it occurred during finals week when students do not have regular classes); however, it seems just as plausible that increased study time would increase confidence in the high performers, and this was clearly not the case.

In conclusion, as Hacker et al. (2000; also see Pierce & Smith, 2001) note, miscalibration of postdiction accuracy has significant implications for student learning; poor accuracy makes it more likely that the student will not be able to effectively allocate future study effort or adapt subsequent task performance. The present study provided new evidence that low performing students suffer from functional overconfidence in their postdiction judgments of performance, but that they also seem to generally be aware that their metacomprehension is not well calibrated. As such our data lend more credence to the hypothesis that the "unskilled are aware" (Miller & Geraci, 2011) rather than the hypothesis that the "unskilled are unaware" (e.g., Dunning et al., 2003; Ehrlinger et al., 2008). Given that low performing students seem fairly aware of their own metacomprehension deficits both before and after testing, but that their confidence may improve under certain conditions, future research should include interventions which focus on (a) increasing awareness of the connection of confidence to actual performance, and (b) requiring students to practice reflecting on past performance more, rather than only being concerned with predicting upcoming performance.

References

Bol, L., Hacker, D.J., O'Shea, P.A., & Allen, D. (2005). The influence of overt practice, achievement level, and explanatory style on calibration accuracy and performance. *Journal Of Experimental Education*, *73*(4), 269-290. doi:10.3200/JEXE.73.4.269-290

Dunlosky, J., Hertzog, C., Kennedy, M., & Thiede, K. (2005). The self-monitoring approach for effective learning. *International Journal of Cognitive Technology, 10,* 4-11.

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. doi:10.1111/j.1467-8721.2007.00509.x

Dunlosky, J., Rawson, K. A., & Middleton, E. L. (2005). What constrains the accuracy of metacomprehension judgments? Testing the transfer-appropriate-monitoring and accessibility hypotheses. *Journal Of Memory And Language*, *52*(4), 551-565. doi:10.1016/j.jml.2005.01.011

Dunlosky, J., Serra, M. J., Matvey, G., & Rawson, K. A. (2005). Second-order judgments about judgments of learning. *Journal Of General Psychology*, *132*(4), 335-346. doi:10.3200/GENP.132.4.335-346

Dunning, D., Johnson, K., Ehrlinger, J., & Kruger, J. (2003). Why people fail to recognize their own incompetence. *Current Directions In Psychological Science*, *12*(3), 83-87. doi:10.1111/1467-8721.01235

Ehrlinger, J., Johnson, K., Banner, M., Dunning, D., & Kruger, J. (2008). Why the unskilled are unaware: Further explorations of (absent) self-insight among the incompetent. *Organizational Behavior And Human Decision Processes*, *105*(1), 98-121. doi:10.1016/j.obhdp.2007.05.002

Glenberg, A. M., & Epstein, W. (1985). Calibration of comprehension. *Journal Of Experimental Psychology: Learning, Memory, And Cognition*, *11*(4), 702-718. doi:10.1037/0278-7393.11.1-4.702

Hacker, D. J., Bol, L., & Bahbahani, K. (2008). Explaining calibration accuracy in classroom contexts: The effects of incentives, reflection, and explanatory style. *Metacognition And Learning*, *3*(2), 101-121. doi:10.1007/s11409-008-9021-5

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. doi:10.1037/0022-0663.92.1.160

Hacker, D. J., Bol, L., & Keener, M. C. (2008). Metacognition in education: A focus on calibration. In J. Dunlosky, R. A. Bjork (Eds.), *Handbook of metamemory and memory* (pp. 429-455). New York, NY US: Psychology Press.

Hertzog, C., Dixon, R. A., & Hultsch, D. F. (1990). Relationships between metamemory, memory predictions, and memory task performance in adults. *Psychology And Aging, 5*(2), 215-227. doi:10.1037/0882-7974.5.2.215

Kelemen, W. L., Winningham, R. G., & Weaver, C. (2007). Repeated testing sessions and scholastic aptitude in college students' metacognitive accuracy. *European Journal Of Cognitive Psychology, 19*(4-5), 689-717. doi:10.1080/09541440701326170

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. doi:10.1037/0096-3445.126.4.349

Krueger, J., & Mueller, R. A. (2002). Unskilled, unaware, or both? The better-than-average heuristic and statistical regression predict errors in estimates of own performance. *Journal Of Personality And Social Psychology, 82*(2), 180-188. doi:10.1037/0022-3514.82.2.180

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. doi:10.1037/0022-3514.77.6.1121

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). New York, NY US: Cambridge University Press. doi:10.1017/CBO9780511489976.004

Maki, R. H., Shields, M., Wheeler, A., & Zacchilli, T. (2005). Individual differences in absolute and relative metacomprehension accuracy. *Journal Of Educational Psychology, 97*(4), 723-731. doi:10.1037/0022-0663.97.4.723

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. doi:10.1037/a0021802

Nietfeld, J. L., Cao, L., & Osborne, J. W. (2005). Metacognitive monitoring accuracy and student performance in the postsecondary classroom. *Journal Of Experimental Education, 74*(1), 7-28.

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

Stine-Morrow, E. L., Gagne, D. D., Morrow, D. G., & DeWall, B. (2004). Age differences in rereading. *Memory & Cognition*, *32*(5), 696-710. doi:10.3758/BF03195860

Thiede, K. W., Anderson, M. M., & Therriault, D. (2003). Accuracy of metacognitive monitoring affects learning of texts. *Journal Of Educational Psychology*, *95*(1), 66-73. doi:10.1037/0022-0663.95.1.66

Footnotes

[1] Clearly these are postdictions and not predictions, but because most students are not familiar with that term, we used the term prediction when asking them to make their judgments of past performance and confidence judgments.

[2] We examined the data from these 21 individuals and did not find any evidence that students who omitted answering either or both of the metacomprehension questions were systematically different in exam performance; four would have been in quartile four (best performing), seven in quartile three, four in quartile two, and six in quartile one.

[3] Instead of using overall average exam grade, one could form the quartiles by some other objective measure of performance, such as final course grade at the end of the seemster. In this course, exams were 60% of the course grade, and other assessments comprised 40%. We examined that possibility using a series of mixed ANOVAs, and the data showed a similar pattern of results to the results we reported using overall average exam grade. That is, for absolute accuracy of postdictions, low performing students did not improve across exams, whereas high performing students did (particularly from Exam 1 to Exam 2); for subjective confidence judgments, differences in confidence tended to be diminished across multiple exams.