*Assessment & Evaluation in Higher Education*
*Vol. 32, No. 2, April 2007, pp. 89–105*

# Discrepant performance on multiple-choice and short answer assessments and the relation of performance to general scholastic aptitude

April Bleske-Rechek[a]*, Nicole Zeug[a] and Rose Mary Webb[b]
[a]*University of Wisconsin-Eau Claire, USA;* [b]*Appalachian State University, USA*

We conducted correlational and performance discrepancy analyses on exam and achievement data taken from students in three psychology courses. Across courses, the same findings emerged. First, only a small fraction of students consistently performed more strongly on one type of assessment (e.g., multiple-choice) than on another (e.g., short answer). Second, students' multiple-choice performance, above and beyond their short answer performance, accounted for variation in students' standing on achievement measures unrelated to psychology (including high school class standing, American College Test score, and college grade point average). In contrast, students' short answer performance, above and beyond their multiple-choice performance, did *not* account for variation in students' standing on those achievement measures. Our findings support the continued use of multiple-choice items to assess student learning.

## Introduction

College instructors face the challenge of creating course exams that are valid and fair, yet efficient, assessments of student learning. Cutbacks in resources for post-secondary education in recent years have heightened this challenge at many higher education institutions by generating larger class sizes and higher teaching loads. To maintain their productivity and efficiency, some instructors have come to rely increasingly upon closed-ended (primarily multiple-choice) rather than short answer or essay items. Such reliance raises the question of whether the exclusive use of one type of assessment is appropriae.

On one hand, closed-ended assessments demonstrate greater content validity and inter-rater reliability than do open-ended assessments (Newstead & Dennis, 1994),

*Corresponding author. Department of Psychology, University of Wisconsin, Eau Claire, 105 Garfield Avenue, Eau Claire, Wisconsin, 54701, USA. Email: bleskeal@uwec.edu

ISSN 0260-2938 (print)/ISSN 1469-297X (online)/07/020089–17

DOI: 10.1080/02602930600800763

and students' scores on closed-ended and open-ended forms of assessment generally correlate highly (i.e., above .5) for college-level tests (Bridgeman & Lewis, 1994; Bridgeman & Morgan, 1996; Kniveton, 1996). Further, many measurement experts historically have favored the use of closed-ended assessments (e.g., Stanley, 1954). Thus, a heightened reliance on closed-ended assessments may be appropriate.

On the other hand, educators have expressed concern that the exclusive use of multiple-choice assessments may put some students at a disadvantage (Bridgeman & Lewis, 1994). Limited research suggests, for example, that a portion of students who perform poorly (i.e., lowest third of the distribution) on multiple-choice assessments perform quite well (i.e., top third of the distribution) on essay assessments, and these students are as likely as those who demonstrate the opposite pattern (i.e., poor essay performance but high multiple-choice performance) to be successful in other college courses (Bridgeman & Morgan, 1996). Thus, educators have expressed concern that students who perform discrepantly on closed-ended and open-ended forms of assessment may be unintentionally disadvantaged in courses that rely exclusively on closed-ended assessments. Given the frequency with which students comment that they excel at one form of assessment over another, one might assume that such unintentional disadvantage occurs frequently. To our knowledge, this question has not been examined empirically. Thus, the first objective of the current study was to determine the actual frequency with which students in different college courses perform well on one form of assessment (in this case, short answer) but poorly on another (in this case, multiple-choice).

Systematic data on the links between (1) performance on closed-ended and open-ended assessments in current college courses, and (2) performance on unrelated measures of general student aptitude should offer concrete guidance on the costs and benefits associated with relying exclusively on one form of assessment over another. Research thus far has focused on closed-ended items of multiple-choice form (as opposed to true/false, matching, or fill-in-the-blank) and open-ended items of essay form (as opposed to short answer); the findings suggest that course-specific essay assessments do not consistently account for variation in individuals' performance on other achievement tasks, beyond that already accounted for by multiple-choice assessments (Miller, 1999). No study to date, however, has investigated the relation of multiple-choice versus *short answer* assessments with performance on measures of general student aptitude. Thus, the second objective of this study was to investigate the degree to which student performance on multiple-choice and short answer items from specific course exams is correlated with their performance on other, more general, measures of student learning. To this end, students' scores on multiple-choice and short answer subsections of course exams were pitted against one another in their relation to students' performance on other indices of academic achievement and student learning. Assuming that items on a course exam are assessing student learning, they should correlate with students' standing on other, general measures of aptitude and student learning. Indeed, measures of aptitude and achievement, including learning within a given course, differ not in kind, but in (a) the degree to which they maintain ties to curricula, (b) their breadth of coverage, (c) their recency of learning, and (d) their purpose of assessment

(Cleary *et al.*, 1975). Any index of student learning *in a specific course* should correlate with performance, to varying degrees, on other indices of student learning *in general*.

In the current research, we utilized high school standing, American College Test (ACT) scores, and college grade point average (GPA) as correlates of exam performance because they are well-known and frequently utilized measures of general student aptitude (i.e., indicators of academic success). We did not use course-specific criteria, such as a course presentation or an outside observer's assessment of each student's learning in the course, because they are more subjective and are not linked reliably to measures of student aptitude (Linn, 1982; Lubinski & Dawis, 1992)[1].

## Method

### *Participants*

Participants were students enrolled in three different psychology courses taught by the first author during the 2003–4 academic year. All students were fully informed about the study procedure before they were asked to consent to participate. The first sample included 29 male and 71 female students enrolled in General Psychology, an introductory level psychology course. The large majority were freshmen. A total of 101 (of 108 enrolled) students were in attendance on the day the study was described. All but one consented.

The second sample consisted of 13 male and 31 female students enrolled in Research Methods in Psychology, a course for students pursuing a major in psychology (none had been in the first sample). The majority of students were juniors. All 44 students enrolled in the course were in attendance on the day the study was described, and all students consented.

The third sample consisted of 5 male and 21 female students enrolled in Evolutionary Psychology, a course designed for advanced students (none had been in the first or second sample). The majority of participants were juniors. Twenty-seven of 28 students were in attendance on the day the study was described, and all but one consented.

### *Instruments*

As part of the study, the instructor designed each course exam to include both multiple-choice and short answer items. The exact split differed by exam and course; thus, Table 1 shows, for each course, the number of exams included in analyses, the week of the academic semester that each exam was taken, the number of multiple-choice and short answer questions on each exam, the percent of total points on each exam that was multiple-choice and the percent that was short answer, the percent of total course points that the exams together contributed toward students' final course grade, and a list of the other activities and assignments that contributed to students' final course grade. As shown in Table 1, the specific structure of each course varied.

On all exams, the closed-ended questions were all four-option multiple-choice in nature. The multiple-choice questions comprised the first section of the exam and the

Table 1. Exam and course specifics

| | General Psychology | Research Methods | Evolutionary Psychology |
|---|---|---|---|
| Number of exams | 2 | 2 | 4 |
| Semester weeks when taken | 6, 12 | 11, 16 | 4, 8, 12, 16 |
| Number of MC items on each exam | 35 | 38 (Exam 1), 41 (Exam 2) | 30 |
| Percent of MC points on each exam | 70% | 57% (Exam 1), 68% (Exam 2) | 75% |
| Number of SA items on each exam | 2 | 2 | 2 |
| Percent of SA points on each exam | 30% | 43% (Exam 1), 32% (Exam 2) | 25% |
| Percent that exams contribute toward final grade | 50% | 23% | 80% |
| Other points toward final grade | Reading quizzes, closed-ended final | Research report and presentation, data analysis exercises, quizzes, article analyses | Three short papers |

short answer questions the second; however, both sections of each exam tested the same content areas, and students completed the exam sections in whichever order they preferred. The first author created test items to cover as many topics as possible from exam study guides and daily lists of course objectives; in this process, a few multiple-choice questions were inspired by test banks but modified. Table 2 displays one sample multiple-choice question and one sample short answer question from each course, and the number of points allotted to each multiple-choice question and the subcomponents of each short answer question.

Both the multiple-choice and short answer sections were limited in reliability, due to a limited number of test items and the possibility of student guessing (which existed for particular short answer items in addition to the multiple-choice items) (Burton & Miller, 1999; Burton, 2001). Students were not penalized for guessing, and were instructed to answer all questions. All students completed all questions. On each exam, there were fewer short answer questions than multiple-choice questions. In an attempt to prevent lower reliability in the short answer sections than in the multiple-choice sections, as is frequently observed in testing, each short answer question consisted of multiple, smaller parts. Each short answer question, in total, required a substantial response (on average, one page of writing) from students. These short answer questions, like the multiple-choice questions, were designed to test students' retention, comprehension, and application of course material. Specifically, questions required students to summarize some unit of class material (*retention*), classify information or reorganize facts (*comprehension*), or generate illustrations or examples (*application*). Not all short answer items were as open-ended as one would ideally have—a couple called for one-word answers or were of fill-in-the-blank format, and thus were more closed-ended in nature (see Table 2)—but very few points were allotted to those items.

The short answer exam sections for the Research Methods course were different from those in the other two courses and thus we explain them in detail here. Because the class material for the first exam under analysis involved experimental designs,

Table 2. Sample multiple-choice and short answer exam items (with point allotments), by course

*General Psychology*

Multiple-Choice

2 pts. Behind Daddy's back, but in front of my two-year-old son, I fill a "Bob-The-Builder Fruit Snacks" box full of crackers. Then I ask my son what Daddy will expect to find in the box when he comes back in the room. This task is known as the_________test, and I can expect my son's answer to be _________.

a. container test; crackers
b. displacement test; crackers
c. container test; fruit snacks
d. displacement test; fruit snacks

Short Answer

Use this page to complete (a) thru (e) below.

2 pts. (a) Bowlby theorized that the attachment behavioral system evolved to serve several functions. Describe what it means to use the primary caregiver as a secure base.

3 pts. (b) Describe the steps of the Strange Situation, a method used by researchers to assess the bond shared between an infant and his/her primary caregiver.

4 pts. (c) List the four different attachment styles generally displayed by infants during the Strange Situation, and for each one identify the primary characteristics of the infant's behavior during the Strange Situation.

4 pts. (d) Is the attachment bond you develop during infancy with you for life? That is, are infant attachment styles destiny? State Yes or No; then, support your answer by describing the findings from a longitudinal study of people's attachment styles from age 1 to age 18.

2 pts. (e) In thinking about day care and attachment, what two things about day care have been found to NOT predict whether a child is securely or insecurely attached? List them. In contrast, what one parental variable DOES predict whether a child will be securely or insecurely attached (regardless of whether the child is in daycare or not)?

*Research Methods*

Multiple Choice

2 pts. Suppose a clinician is interested in finding out whether Alcoholics Anonymous (AA) will help his clients' alcoholism. He does an initial assessment of one of his client's drinking behavior, and then begins this client on a two-month intervention of AA meetings three times per week. At the end of two months, he assesses his client and notes that his client's drinking has decreased drastically. At this point, the clinician knows he cannot withdraw the treatment. To really know if AA is effective for his clients, he must pursue not a _________ design, but rather a _________ design.

a. multiple-baseline across situations; ABAB
b. ABAB; multiple-baseline across situations
c. multiple-baseline across individuals; multiple-baseline across behaviors
d. ABAB; multiple-baseline across individuals

Table 2. *(Continued)*

Short Answer

Read the following description of a research study. Then, answer the questions that follow. Professor K is interested in testing whether a man's physical attractiveness and apparent financial capacity interact to influence women's level of sexual attraction to him. Professor K takes photographs of various men who have been previously rated as unattractive, attractive, or very attractive. He then duplicates the photographs so that each unattractive man is pictured once in a McDonald's uniform and once in a business suit; then, he does the same for each attractive man and each very attractive man. Professor K randomly assigns his 300 female participants to one of three conditions. Participants in the first condition rate each *unattractive* model twice – when he is pictured in a McDonald's uniform and when he is pictured in a business suit. Participants in the second condition rate each *attractive* model twice – when he is pictured in a McDonald's uniform and when he is pictured in a business suit. Participants in the third condition rate each *very attractive* model twice – when he is pictured in a McDonald's uniform and when he is pictured in a business suit. Participants report their level of sexual attraction to the man in each photograph on a 10-point scale (1 = Not at all sexually attracted, to 10 = Highly sexually attracted).

2 pts. 1. This study is a _________x _________ (choose one) within-subjects/between-subjects/ mixed-subjects design.

1 pt. 2a. State the first factor (variable) and the number of levels it has:

1 pt. 2b. Is this an independent variable or individual-differences variable?

1 pt. 2c. Is this a between-subjects or within-subjects variable?

1 pt. 3a. State the second factor (variable) and the number of levels it has:

1 pt. 3b. Is this an independent variable or individual-differences variable?

1 pt. 3c. Is this a between-subjects or within-subjects variable?

2 pts. 4. State the dependent variable and how it is operationalized:

2 pts. 5. What general technique do you suppose Professor K is using to make sure that the order of presentation of conditions varies from participant to participant?

Professor K finds a *main effect of financial capacity*: Across levels of attractiveness, the women reported higher levels of sexual attraction to men who were high in apparent financial capacity (i.e., dressed in business suits). Professor K also finds an *interaction*: For men low in apparent financial capacity, being physically attractive didn't help at all to win women's hearts. For men high in apparent financial capacity, however, increasing levels of physical attractiveness were associated with increasingly higher levels of reported sexual attraction.

7 pts. 6a. First, draft a table (similar to that shown for the previous study) of imaginary group mean ratings that portray these two significant effects (as well as a *non*-significant main effect of attractiveness). For your main effect, you should consider a difference of 2 or more (on the rating scale of 1 to 10) between marginal means to be statistically significant.

4 pts. 6b. Second, sketch out the results in a clustered bar graph, with "Apparent Financial Capacity" on the X-axis. Label your axes, and make sure your Y-axis scale gives the full range of possible means.

*Evolutionary Psychology*

Multiple-Choice

2 pts. How do we know that females' advantage in location memory is not just a side effect of females being more attentive than males to their surroundings?

a. the female advantage persists in directed learning situations
b. the female advantage persists in incidental learning situations
c. females have more of an advantage in object memory than in location memory
d. males have an advantage in object memory, whereas females have an advantage in location memory

Table 2. *(Continued)*

Short Answer

Develop an essay that answers the following questions:

2 pts. (a) *Why* should dads be sensitive to cues of their paternity, such as a child's resemblance to them?

4 pts. (b) Describe findings from one study that demonstrate that men are, indeed, sensitive to (and influenced by) whether a child resembles them.

4 pts. (c) Do children, on average, *actually* resemble their fathers more than their mothers? Explain.

both sections of this exam focused on experimental design and students' comprehension and application of the concepts of main effects and interactions. The second short answer question, shown in Table 2, was unique conceptually because it asked students to work 'backwards' through a problem (the study's pattern of findings were given; students were asked to generate the numbers that would implicate that pattern), which they had not been asked to do in class. It also involved a two factor design, with one factor having *three* levels rather than the customary two, so students were expected to extend their knowledge to a more complex example than they previously had been exposed to in class. Exam 2 was similar in the extra conceptual demand placed on the short answer questions. The short answer section again included one item that went beyond students' explicit knowledge by asking them to discuss a *three* factor study and extrapolate their knowledge of two-way interactions to describe a three-way interaction.

The first author and a teaching apprentice scored all multiple-choice items by hand. The first author, blind to students' identity and score on the multiple-choice section (short answer sections began on a page separate from multiple-choice sections), scored all short answer items.

### *Procedure*

After semester grades were posted, researchers found out more about each student, with his or her consent, from the university registrar. For the first sample, researchers accessed term GPA, ACT score, and high school standing (in percentile); for the second and third samples, researchers accessed cumulative GPA as well.

## Results

### *Descriptive statistics*

Table 3 displays descriptive statistics on the relevant variables for each sample. Due to a small number of male students across the three samples, we combined the sexes for all analyses. All exam scores, as percentages, reflect (number correct)/(number possible). Overall, means centered around 75 per cent. Substantial variation existed for all variables, with neither floor nor ceiling effects a major concern.

Table 3. Descriptive statistics for correlated variables

| | General Psychology<br>*M* (*SD*) | Research Methods<br>*M* (*SD*) | Evolutionary Psyc<br>*M* (*SD*) |
|---|---|---|---|
| *Exam Variables (of 100%)* | | | |
| Multiple-Choice Average | 70.74 (9.94) | 77.69 (9.90) | 77.05 (7.35) |
| Short Answer Average | 65.40 (14.90) | 83.12 (9.08) | 75.14 (10.96) |
| *General Achievement Measures* | | | |
| High School Percentile | 76.67 (13.03) | 70.50 (19.18) | 63.12 (24.95) |
| ACT Composite Score (of possible 36) | 23.62 (2.44) | 23.77 (3.34) | 23.74 (3.29) |
| University Term GPA (of possible 4.00) | 3.01 (.55) | 3.17 (.53) | 3.06 (.52) |
| University Cumulative GPA (of possible 4.00) | (NA) | 3.01 (.41) | 2.99 (.37) |

*Note*: *M* = Mean; *SD* = Standard Deviation; ACT = American College Test. Please contact the authors for descriptive statistics by exam. Due to missing ACT scores, sample sizes ranged from 90 to 100 for General Psychology, 35 to 44 for Research Methods, and 23 to 26 for Evolutionary Psychology.

### *Reliability analyses and consistency of student performance*

*General Psychology.* Reliability analyses (Cronbach, 1951) were conducted to investigate, in this context, the inter-exam consistency for each assessment type. The resulting alpha reliability coefficient for the two multiple-choice sections was .75, and for the two short answer sections, .46.

Table 4 (above main diagonal) shows that students' scores on the two forms of assessment within each exam were correlated; moreover, students' performance on the first exam, on one form of assessment (e.g., multiple-choice), significantly correlated with their performance on the second exam, on the other form of assessment. The correlation coefficients were consistently large, with exception to two (.29 and .31), both of which involved the short answer section of Exam 1. It is possible that students' overall solid performance on the second short answer question ($M$ = 84.50) accounts for this lower reliability of Exam 1's short answer component. Although students' performance on the first short answer question of Exam 1 correlated positively with their multiple-choice score on Exam 1, $r(100) = .31$, $p < .01$, students' performance on the second short answer question of Exam 1 did not correlate with their multiple-choice score on Exam 1, $r(100) = .17$, $p = .10$.

*Research Methods.* The alpha reliability coefficient for the two multiple-choice sections was .79, and for the two short answer sections, .74.

Table 4 (below main diagonal) shows that students' scores on the two forms of assessment within each exam were correlated; moreover, students' performance on

Table 4. Consistency of student performance across exams and across assessment type: correlations between exam subtests

| | MC Exam 1 | MC Exam 2 | MC Average | SA Exam 1 | SA Exam 2 | SA Average |
|---|---|---|---|---|---|---|
| MC Exam 1 | (1.00) | .60 | .88 | *.31* | .59 | .57 |
| MC Exam 2 | .67 | (1.00) | .90 | .42 | .59 | .64 |
| MC Average | .93 | .89 | (1.00) | .41 | .66 | .68 |
| SA Exam 1 | .62 | .50 | .62 | (1.00) | *.29* | .76 |
| SA Exam 2 | .59 | .58 | .64 | .62 | (1.00) | .84 |
| SA Average | .66 | .61 | .70 | .86 | .93 | (1.00) |

*Note*:. Correlations above the main diagonal are from the General Psychology sample; correlations below the main diagonal are from the Research Methods sample. All correlations are significant at $p < .001$, except for those in italics (.29 and 31), which were significant at $p < .01$. MC = Multiple-Choice; SA = Short Answer. Correlations between MC Exam 1 and MC Average, between MC Exam 2 and MC Average, between SA Exam 1 and SA Average, and between SA Exam 2 and SA Average are spuriously high because they are part-whole correlations.

the first exam, on one form of assessment (e.g., multiple-choice), significantly correlated with their performance on the second exam, on the other form of assessment. Correlation coefficients were consistently large.

*Evolutionary Psychology.* The alpha reliability coefficient for the multiple-choice sections was .66, and for the short answer, .65. Due to the small sample size of this class ($N = 26$) relative to that of the other two classes, and the relatively low number of closed-ended exam questions on the exams in this course relative to that of the other courses (see Table 1), we created a 'multiple-choice' composite variable and a 'short answer' composite variable for subsequent analyses of performance discrepancies. The composite variables were created by taking the students' average performance across the four exams. The two forms of assessment demonstrated comparable consistency: students' performance on each multiple-choice section correlated strongly with their overall short answer performance (average $r = .49$, all $p$s $< .001$), and students' performance on each short answer section correlated strongly with their overall multiple-choice performance (average $r = .48$, all $p$s $< .001$).

### *Performance discrepancies*

*General Psychology.* We examined discrepancies in performance on the two forms of assessment by computing for each student, for each exam, a discrepancy score equal to their multiple-choice score (percent) minus their short answer score (percent). Thus, an above-zero discrepancy score indicated better performance on the multiple-choice section, and a below-zero discrepancy score indicated better performance on the short answer section. Across students, the mean discrepancy

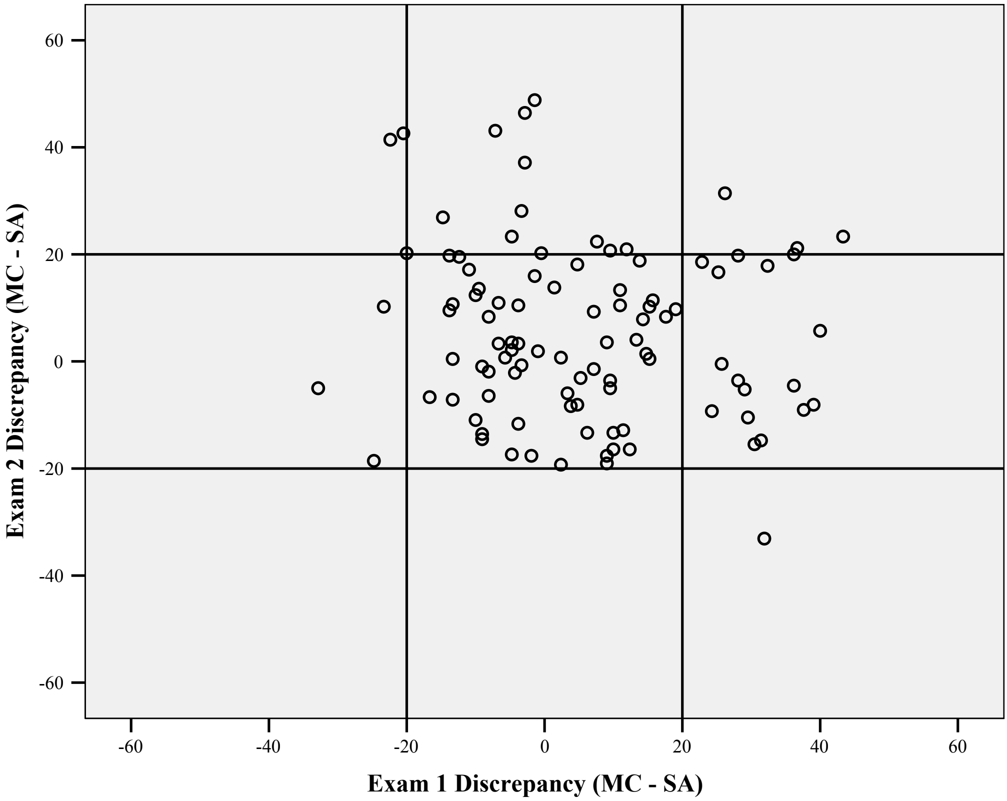

Figure 1. General Psychology sample: simple percentage discrepancy scores on Exam 1 plotted against simple percentage discrepancy scores on Exam 2 (r(99) = –.14, ns)

score for Exam 1 was 5.34 (*SD* = 16.85; range –32.86 to +43.33), and for Exam 2 it was 5.25 (*SD* = 16.50; range –33.10 to +48.81). As displayed in Figure 1, students' discrepancy scores on Exam 1 were not related to their discrepancy scores on Exam 2, *r*(99) = –.13, *p* = .18; that is, students were not consistently favored by one form of assessment over another.

Large performance discrepancies (20 percentage points or more) are displayed in the four outer boxes of Figure 1. Four students of 99 (4%) achieved on both exams a multiple-choice score that exceeded their short answer score by 20 percentage points or more (e.g., a score of 90% on the multiple-choice section versus a 70% on the short answer section). Only one student achieved on both exams a multiple-choice score that exceeded their short answer score by more than 25 percentage points. These discrepancies themselves may be chance aberrations: another four students achieved on one exam a discrepancy score of +20 or more (multiple-choice favored) but on the other exam a discrepancy of –20 or more (short answer favored). No student achieved on both exams a short answer score that exceeded their multiple-choice score by 20 percentage points or more.

Following Bridgeman and Morgan (1996), we also split students into top and bottom 'thirds' on the basis of their performance on the multiple-choice and short answer sections. Not one student in General Psychology scored in the top third of one type of assessment and the bottom third of the other type across both mid-semester exams.

*Research Methods.* We again examined discrepancies in performance on the two forms of assessment by computing for each student, for each exam, a discrepancy score. Across students, the mean discrepancy score for Exam 1 was –9.24 (*SD* = 9.56; range –27.05 to +13.34), and for Exam 2 it was –1.61 (*SD*= 9.92; range –26.06 to +16.75). As displayed in Figure 2, students' discrepancy scores on Exam 1 were not related to their discrepancy scores on Exam 2, *r*(44) = .16, *p* = .30; again, students were not consistently favored by one form of assessment over another.

Large performance discrepancies are displayed in the four outer boxes of Figure 2. Two students of 44 (4.5%) achieved on both exams a short answer score that exceeded their multiple-choice score by 20 percentage points or more (a grade of 'A' on short answer versus a grade of 'C' on multiple-choice). Not one student achieved

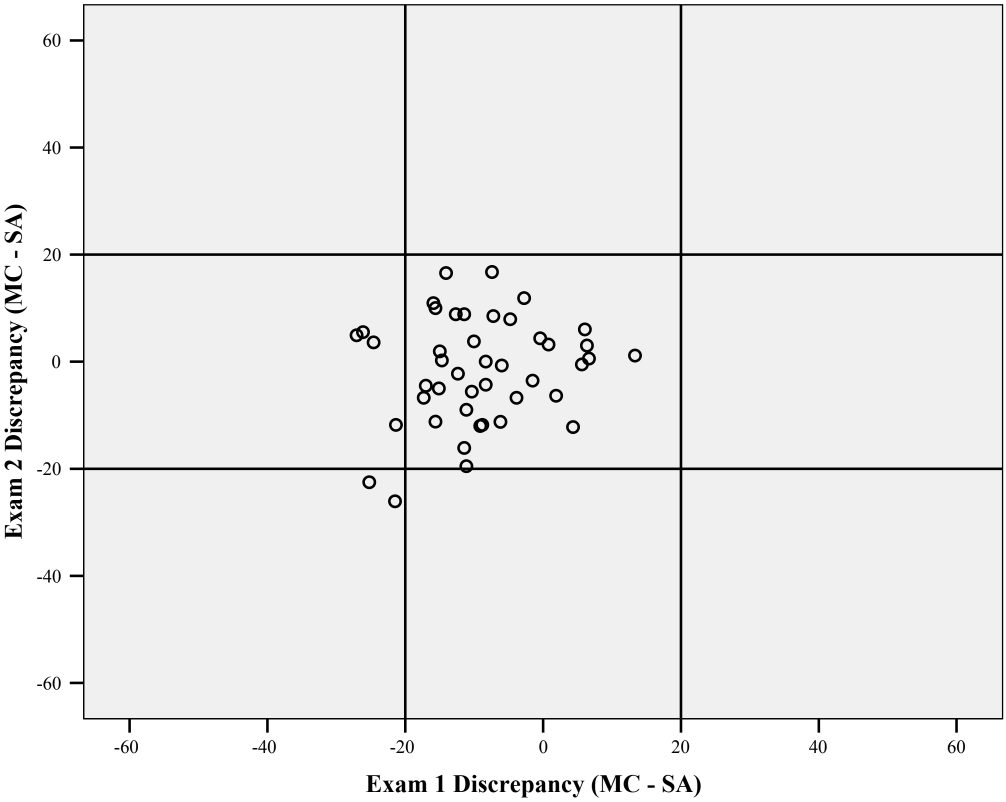

Figure 2. Research Methods sample: simple percentage discrepancy scores on Exam 1 plotted against simple percentage discrepancy scores on Exam 2 (r(44) = .16, ns)

on both exams a short answer score that exceeded their multiple-choice score by more than 25 percentage points. Not one student achieved on both exams a multiple-choice score that exceeded their short answer score by 20 percentage points or more. Further, the top and bottom thirds split revealed that no student scored in the top third of the class on one type of assessment and the bottom third on the other type across both mid-semester exams.

*Evolutionary Psychology.* We examined discrepancies in performance on the two forms of assessment by computing for each student, for each exam, a discrepancy score; we also computed a discrepancy score to represent average multiple-choice performance compared to average short answer performance. Not one student in Evolutionary Psychology exhibited a composite discrepancy of 20 (or even 15) percentage points (Mean = 1.91, SD = 7.90; range −13.54 to +14.38). On an exam-by-exam basis, only two students twice achieved a multiple-choice score that exceeded their short answer score by 20 or more percentage points. Only one student achieved *three* times a multiple-choice score that exceeded their short answer score by more than 20 percentage points. As with the General Psychology sample, it is possible that these are chance aberrations: three students achieved on one exam a discrepancy score of +20 or more (multiple-choice advantage) and on another exam a discrepancy score of −20 or more. No student achieved more than once a short answer score in 20-point excess of their multiple-choice score. Correlations between the individual exam discrepancy scores ranged from −.25 to .09, all *p*s > .21. As with the other samples, students were not consistently favored by one form of assessment over another.

As was done with the other samples, students were also split into top and bottom 'thirds' on the basis of their composite performance on the multiple-choice and short answer sections. Averaged across the four exams, only one student of 26 performed in the top third of the class on the short answer sections despite poor performance overall on the multiple-choice sections.

### *Relation of exam performance to general scholastic aptitude*

*General Psychology.* The top panel of Table 5 displays the results of partial correlational analyses that were conducted with General Psychology exam scores to test the relationship between general scholastic aptitude and performance on each form of assessment while holding constant performance on the other form of assessment. As shown in the table, all zero-order bivariate correlations were significant. Students' multiple-choice performance continued to be correlated with their general scholastic aptitude after controlling for their short answer performance; however, students' short answer performance did not correlate with general scholastic aptitude after controlling for their multiple-choice performance.

*Research Methods.* The middle panel of Table 5 displays the results of these analyses for the Research Methods sample. The findings replicated those from the

Table 5. Zero-order and partial correlations between (1) MC and SA exam scores and (2) measures of general student aptitude

| | *Multiple-Choice (MC) Score* | | *Short Answer (SA) Score* | |
|---|---|---|---|---|
| | Zero-order *r* | Partial *r* (MC Controlling for SA) | Zero-order *r* | Partial *r* (SA Controlling for MC) |
| *General Psychology* | | | | |
| High School Percentile | .47*** | .29** | .41*** | .14 |
| ACT Score | .30** | .20* | .22* | .02 |
| Term GPA | .69*** | .50*** | .60*** | .29** |
| *Research Methods* | | | | |
| High School Percentile | .49** | .42** | .29 | −.08 |
| ACT Score | .58*** | .34* | .54** | .22 |
| Term GPA | .68*** | .56*** | .47** | −.02 |
| Cumulative GPA | .75*** | .60*** | .57*** | .09 |
| *Evolutionary Psychology* | | | | |
| High School Percentile | .17 | .17 | .07 | −.08 |
| ACT Score | .29 | .13 | .29 | .12 |
| Term GPA | .67*** | .49* | .54** | .15 |
| Cumulative GPA | .59** | .41* | .48* | .12 |

*Note*: * $p < .05$, ** $p < .01$, *** $p < .001$. MC = Multiple Choice; SA = Short Answer. MC and SA represent students' average MC and average SA performance, respectively, across exams. For General Psychology, *df* = 92–94; for Research Methods, *df* = 33–42; for Evolutionary Psychology, *df* = 21–24

General Psychology sample. For all variables of interest, students' multiple-choice performance continued to be correlated with their general scholastic aptitude after controlling for short answer performance; however, short answer performance did not correlate with general scholastic aptitude after controlling for students' multiple-choice performance.

*Evolutionary Psychology.* The bottom panel of Table 5 displays the results of these analyses for the evolutionary psychology sample. In two of four cases, students' multiple-choice performance continued to be correlated with their general scholastic aptitude after controlling for short answer performance; in no case did short answer performance correlate with general scholastic aptitude after controlling for students' multiple-choice performance.

## Discussion

The significance of this investigation is twofold. First, this investigation is the first to document empirically the frequency with which students in a psychology course perform well on one form of assessment but poorly on another. Utilizing percentage difference scores and top-third and bottom-third splits, we found that, contrary to popular notions and anecdotal reports, students infrequently perform discrepantly on short answer and multiple-choice assessments. Second, this study is the first to assess the links between scores on different forms of assessment and scores on common measures of general student aptitude and achievement. In three separate psychology courses, we found that multiple-choice assessments of student knowledge demonstrated links with measures of general student aptitude, even after controlling for short answer assessments of student knowledge. Shorter answer assessments did not show similar independent links with scholastic aptitude.

### *Performance discrepancies*

Given comments instructors sometimes hear from students, such as, 'I'm just really poor at multiple-choice tests,' or 'I do fine on multiple-choice questions; it's the essays that get me,' one might expect performance discrepancies to be somewhat frequent. In our samples, students rarely performed 20 percentage points or better on one form of assessment than on another (for example, 90% on one section but 70% on another) across more than one exam. Although some instructors may have realized through their teaching experience that performance discrepancies are not as common as students may believe, our systematic examination provides more compelling, data-driven evidence of the prevalence of performance discrepancies, at least for students in psychology. The data compiled here provide instructors with an empirical response to students with performance discrepancy concerns. Regardless of the practical utility of the finding, it parallels that of research on measures of general mental ability: Forms of assessment may vary widely (e.g., a traditional IQ test and a Hicks Paradigm decision-making task), but to the extent that they all draw on the same construct, participants' scores on the assessments correlate positively (Spearman, 1927; Jensen, 1998; Deary, 2001).

### *Exam performance and general scholastic aptitude*

If a course exam is, in fact, an assessment of student learning, then students' performance on it should correlate with past and current measures of students' achievement (or aptitude, ability: see Cronbach, 1990; Lubinski & Dawis, 1992). Although high school percentile, ACT score, and college GPA all serve, to varying degrees, as indicators of student achievement, each also has specific variance associated with it. Our analyses showed that, for students in three different psychology courses, multiple-choice test questions consistently exceeded short answer test questions in their independent links to measures of student achievement. Particularly compelling is that multiple-choice performance explained substantial variation in semester GPA and cumulative GPA, measures of student learning that presumably entail varied forms of assessment (e.g., research papers, presentations, lab assignments) across multiple disciplines.

The results from the Research Methods sample arguably are more compelling than those of the other samples. The course activities of most Research Methods courses, which include data analysis assignments, critical thinking exercises, report writing, and presentations, implicate open-ended assessments as more appropriate than closed-ended assessments. Again, however, the multiple-choice assessments fared better than the short answer assessments across all tests conducted.

### *Reliability*

Our analyses imply—albeit tentatively—that instructors should be more wary of exclusive reliance on short answer assessments in psychology than of exclusive reliance on multiple-choice assessments in psychology. Psychometricians issued this warning in the past (Stanley, 1954), but for a different reason—lower reliability of open-ended assessments as compared to closed-ended assessments. However, lower reliability is not a likely explanation of the pattern of findings from the current investigation because in two of the three samples, alpha reliability coefficients were as high for short answer assessments as for multiple-choice assessments. Moreover, student performance on short answer sections correlated strongly with student performance on the multiple-choice sections. Because our open-ended items were short answer rather than essay (which sometimes require the formulation and development of a single argument), they actually may have been more reliable and consistently scored than open-ended items used in past investigations and thus may have provided a more conservative test of the unique relation between performance on course-specific closed-ended exams and scholastic achievement.

### *Limitations*

The current findings, however, should be interpreted with caution. First, all exams under analysis were likely limited in reliability due to the number of test items and the possibility of student guessing. Second, the exams were taken only from psychology courses and from three courses taught by the same instructor. Certain elements of the instructor's style of teaching and exam writing may have been common to all three courses and thus may be confounded with the common results across the three courses. That said, it is noteworthy that, as displayed in Table 1, each course's structure was quite different, thus providing the advantage of having three different sources of potential error operating. Future research could investigate performance discrepancies and correlates of different forms of assessment in courses from other disciplines taught by other instructors.

## Conclusion

Our experience is that it is not uncommon for instructors to include a few short answer items on an exam because they or their students perceive that open-ended items provide students with more opportunity to demonstrate what they have learned. Our findings suggest that instructors may gain relatively little, objectively, by including short answer items in addition to multiple-choice items, but much by including multiple-choice items in addition to short answer items. We hope that this investigation will spur further research on the issues surrounding multiple-choice and short answer assessments in other college level courses.

## Acknowledgements

The authors thank Allen Keniston and two anonymous reviewers for comments on earlier versions of this manuscript.

## Notes

1. In one of the three classes, Research Methods, many points were generated from activities besides the exams. Students completed, in pairs, a final research report of 15–20 pages in length. Students also individually completed nine statistical assignments, each of which included drafting a section of a research report in APA format. Finally, students individually completed eight article analyses. We did not use these variables as criteria in our analyses because there were no parallel assignments in the other courses. Across multiple semesters of Research Methods, however, correlations between individual exam scores and final research report scores have been moderate (*r*s .30–.37, *p*s < .05); correlations between individual exam scores and average statistical assignment scores have been high (*r*s .60–.70, *p*s < .001); and correlations between individual exams scores and average article analysis scores have been high (*r*s .59–.65, *p*s < .001). In other words, students' exam performance is highly linked with their performance on other indices of their learning in the course. The lower correlations between exam performance and final report performance are likely due to unreliability produced by having final reports completed by students in pairs rather than individually.

## Notes on contributors

April Bleske-Rechek is an assistant professor of psychology at the University of Wisconsin-Eau Claire, with primary research interests in individual differences and personal relationships.

Nicole Zeug is an advanced undergraduate student at the University of Wisconsin-Eau Claire.

Rose Mary Webb is an assistant professor of psychology at Appalachian State University; her primary research interests are in individual differences and talent development.

## References

Bridgeman, B. & Lewis, C. (1994) The relationship of essay and multiple-choice scores with grades in college courses, *Journal of Educational Measurement,* 31, 37–50.

Bridgeman, B. & Morgan, R. (1996) Success in college for students with discrepancies between performance on multiple-choice and essay tests, *Journal of Educational Psychology,* 88, 333–340.

Burton, R. F. (2001) Quantifying the effects of chance in multiple-choice and true/false tests: question selection and guessing of answers, *Assessment and Evaluation in Higher Education,* 26, 41–50.

Burton, R. F. & Miller, D. J. (1999) Statistical modeling of multiple-choice and true/false tests: ways of considering, and of reducing, the uncertainties attributable to guessing, *Assessment and Evaluation in Higher Education,* 24, 399–411.

Cleary, T. A., Humphreys, L. G., Kendrick, S. A. & Wesman, A. (1975) Educational uses of tests with disadvantaged students, *American Psychologist,* 30, 15–41.

Cronbach, L. J. (1951) Coefficient alpha and the internal structure of tests, *Psychometrika,* 16, 297–334.

Cronbach, L. J. (1990) *Essentials of psychological testing* (4th edn) (New York, Harper & Row).

Deary, I. J. (2001) *Intelligence: a very short introduction* (New York, Oxford University Press).

Jensen, A. R. (1998) *The g factor: the science of mental ability* (Westport, Praeger Publishers).

Kniveton, B. H. (1996) A correlational analysis of multiple-choice and essay assessment measures, *Research in Education,* 56, 73–84.

Linn, R. (1982) Ability testing: individual differences, prediction, and differential prediction, in: A. K Wigdor & W. R. Garner (Eds) *Ability testing: uses, consequences, and controversies* (Washington, DC, National Academy Press), 335–388.

Lubinski, D. & Dawis, R. (1992) Aptitudes, skills, and proficiencies, in: M. D. Dunnette & L. M Hough (Eds) *Handbook of industrial/organizational psychology* (2nd edn) (Palo Alto, Consulting Psychologists Press), 1–59.

Miller, J. (1999) What's wrong at E.T.S.? Insider's view of grading A.P. government essays, *College Teaching,* 47, 2–7.

Newstead, S. & Dennis, I. (1994) The reliability of exam marking in psychology: examiners examined, *Psychologist,* 7, 216–219.

Spearman, C. (1927) *The abilities of man: their nature and measurement* (New York, Macmillan).

Stanley, J. C. (1954) *Measurement in today's schools* (3rd edn) (New York, Prentice-Hall).