Motivation in Low Stakes Testing Conditions:

What’s the Feedback on Feedback?

Donna L. Sundre,
Philip Erb,
Javarro Russell

Symposium paper presentation
The Northeastern Educational Research Association
October 2009
Introduction

One of the most pervasive and insidious problems facing assessment practitioners and measurement experts involves arriving at appropriate interpretations concerning student performances in low stakes testing conditions. If the observed performances of students appear quite high, we may conclude that our students have achieved our learning objectives and effectively demonstrated their learning, or perhaps the test is too easy. Unfortunately, this result is quite uncommon; the interpretive scenario is seldom this clear. In most low stakes testing conditions, the observed performances rarely achieve the levels we desire, and the test users and policy makers have a variety of other quite viable alternative hypotheses to consider: the test is too difficult; our expectations are too high; the students did not learn the material; or perhaps, the students elected not to demonstrate their true abilities. The latter is generally attributed to low student motivation. After concerted effort to: 1) improve the psychometric quality of our instruments; 2) professionalize our data collection procedures; and 3) address other potential confounds, we have spent a great deal of time assessing examinee motivation and discovering ways to improve it. We have learned a lot. This paper describes a recent effort to improve our assessment practice that we thought might, as an additional and intended consequence, positively influence examinee motivation.

Primarily, we have been waiting for our University Student Information System to stabilize to a point where we could provide students the opportunity to review their assessment scores and receive interpretive information. We had previously, and through great effort, provided feedback to students via email; however, the University email vendors changed far too rapidly for us to consistently provide this service. The University recently implemented eCampus, a web-based one-stop location students can easily access for unofficial transcripts,
financial accounting, changing their address information, and many other useful functions. We
knew that eCampus now provided a secure location that students already have experience using.
We have wanted to improve our assessment practice by providing students with feedback in
compliance with the Standards for Educational and Psychological Testing (AERA, APA,
NCME, 1999), and The Program Evaluation Standards (Sanders, 1994). We were pleased when
a long term opportunity finally arrived and wondered... “If we informed examinees at the time of
testing that they would have the opportunity to receive feedback on their scores along with
interpretive information, would it improve their test taking motivation and performance on the
tests?” There has been mention and a few actual studies in the literature of feedback and other
‘incentives’ being used as test taking motivators; but, the results have not been clear.

In general, most incentives have not proven effective in improving student participation
or performance in low-stakes testing conditions. Everything from free pizza, graduation gowns
for seniors, tee shirts, test feedback, and monetary rewards have been attempted. Brown and
Walberg (1993) found that providing motivating instructions produced higher performance, but
the subjects were elementary school students. Trombley (2001) reported serious motivation
issues for college students with most University of Missouri students electing not to participate
in assessment testing despite all of the above incentives offered. Olsen & Wilson (1991)
interviewed over 100 college students who performed well below their predicted score on
assessment tests at Northeast Missouri University (now Truman State University). Their results
suggested that incentives such as gaining free electives for high scores, additional coursework for
low scores, and the promise of feedback might result in higher achievement. Low motivation to
perform well and dislike for standardized tests clearly influenced these individuals.
Several studies have focused on monetary rewards alone with fairly disappointing results. Olsen & Wilson’s (1991) subjects indicated that monetary incentives would have lowered motivation and performance. While no study has yet reported monetary incentives resulting in diminished performance or motivation, the general results do not suggest that monetary incentives would be effective. O’Neil, Sugrue & Baker (1995/1996) found that $1 monetary rewards for each correct answer were effective for 8th grade boys; however, these results did not generalize to 12th graders of either gender. O’Neal, Abedi, Miyoshi & Mastergeorge (2005) more recently attempted another study in which $10 per correct answer was offered to AP high school students. The incentives were not effective in producing higher performance. Clearly, more research is needed.

The incentive of providing feedback to students on their test performance is a “win-win” situation. First of all, it’s the right thing to do. Secondly, students consistently report that they would like to know how they perform on low-stakes tests. V. Wise (2004) reported that the college students in her study indicated they would like to know how they performed on the assessment tests. However, when the promise for test feedback was provided, neither examinee self-reports of test taking motivation nor performance was influenced. S. Wise (in press) recommended more research on what kind of feedback examinees would find most valuable.

Method

Procedures

James Madison University conducts two Assessment Days each year for the purpose of systematically collecting compelling data by which the efficacy of academic programs can be evaluated. Students eligible for assessment are required to participate in these activities. If a
student does not participate, they receive a hold on their registration for the next semester until they have completed their assessment responsibilities. The first Assessment Day of the academic year takes place in August and is an integral part of a four-day required Orientation process for entering first year students. Students are assigned to testing rooms via the last two digits of their JMU ID numbers. The testing rooms are assigned 4-5 assessment tests that include measures of general education (e.g., Quantitative Reasoning, Health and Wellness, the American Experience) and non-cognitive assessment tools (e.g., Belongingness, Achievement Motivation Learning Goals). In other words, students do not take all of the assessment tests. The sampling procedures produce large, representative, and random samples, and inferences can confidently be made about the target population. At the second Assessment Day that occurs in February of the spring semester, the University cancels classes from 8 am to 4 pm. In this way, there are no time or room conflicts for any students. All students who have accumulated 45-70 credit hours, by whatever means (AP, IB, transfer, dual enrollment) receive two emails informing them that they are eligible for spring Assessment Day. Again, students are assigned to testing rooms using the last two digits of their student IDs. Since these numbers do not change, this allows the same students to be assigned the same assessment instruments. In other words, repeated measures are a design feature.

Participants

The primary participants for this study were randomly assigned to one of two rooms, designated as Control and Experimental, for this research project. For each Assessment Day, half of the students are tested in the morning and half are tested in the afternoon. Our Assessment Day participation rates have been quite impressive: Spring 2009 (89%) and Fall 2009 (95%). Of course, following make-up activities, the participation rates are 100%. As with all testing
activities for which the personal consequences for students are minimal (often termed low-stakes), examinee motivation is a potential threat to the validity of any inferences we may wish to make about student learning and growth. This study focused on student performances on two instruments administered on the Spring 2009 Assessment Day and their later access to feedback opportunities.

Secondary study participants included all remaining spring 2009 and entering fall 2008 students who completed the QR SR instruments during their respective Assessment Day session. These individuals were informed via email that their scores and interpretive information were available for review.

Measures

*Natural World-9 (NW-9)*

Both groups of students completed the NW-9. The NW-9 assessment instrument was designed by faculty and assessment specialists at JMU to measure the objectives of the Cluster 3 (CL3), the Natural World segment of JMU’s general education program (Sundre, 2008). To date, there have been nine forms of the Natural World test, designed to assess eight objectives in Cluster 3 (CL3) of James Madison University’s General Education program (see: [http://www.jmu.edu/gened/cluster3.shtml](http://www.jmu.edu/gened/cluster3.shtml)). The current form of the test, NW-9, has 66 multiple-choice items and yields both a Total score (SR) and a Quantitative Reasoning (QR) subscale score. The items are designed to assess process rather than course content specific knowledge. In other words, no item can privilege one particular course over another. The NW-9 is intended to assess a college student’s quantitative and scientific reasoning skills. Reliability estimates are generally in the mid .80s for sophomores and juniors and the mid to high .70s for entering first year students.
Student Opinion Scale (SOS)

Both groups of students completed the SOS (Sundre, 2007; Sundre & Moore, 2002), a measure of examinee motivation. Motivation is operationally defined by two subscales: Importance (i.e. how important the test was to the student) and Effort (i.e. how much effort the student put into test performance). Five items are associated with each subscale. The response scale ranges from 1 “Strongly Disagree” to 5 “Strongly Agree.” Reliability estimates for both subscales are generally in the mid .80s when administered in low-stakes testing conditions. Reliabilities decline dramatically in high-stakes testing contexts.

Procedure

The current study was performed in conjunction with the spring 2009 Assessment Day at JMU. The primary participants (N=534) were assigned to one of two rooms. In addition to several other measures, both groups of students completed the SOS and the NW-9. All tests were given in the same order: 1) a brief non-cognitive assessment on achievement motivation; 2) the NW-9; 3) a cognitive assessment on Society and Politics; and 4) the SOS. Test instructions for the Natural World-9 (NW-9) were manipulated for the two groups of students. Students in the experimental group (N=218) received instructions notifying them that they would have the opportunity to view their NW-9 scores and receive two kinds of interpretive feedback: 1) how they performed in relation to others (NR: normative-referenced); and 2) how they performed in relation to standards set by JMU faculty (CR: criterion-referenced). Appendix A provides the special test instructions this group received. The students in the control group (N=316) received standard test instructions with no mention of feedback opportunities. A primary purpose of the study was to observe whether a very subtle test instruction intervention might result in differences in examinee motivation and/or test performance. These data inform the first stage of
analysis. A secondary purpose of the study was to monitor students’ access and preference for interpretive information about their assessment performances in low-stakes testing conditions. These data inform the second stage of analysis.

Students in the experimental and control conditions completed the NW-9 and the SOS in February, 2009. Later in the spring semester, all students who completed the NW-9 were systematically notified through email that feedback on their NW-9 scores was available through eCampus. A series of emails were sent to different sets of examinees informing them that their scores were available for viewing on eCampus: 1) March 25th: the experimental group; 2) April 8th, the control group; 3) April 15th all other spring 2009 NW-9 test takers; and 4) April 22nd all fall 2008 entering students who completed the NW-9. All emails were sent on a Wednesday during lunch hour. Appendix B provides the letter that was sent to the experimental condition students. This letter was used for the mail merge. Appendix C provides the letter that was slightly modified to inform each of the other student groups who had completed the NW-9 during the 2008-2009 academic year.

It will be recalled that our primary purpose in making the scores available was to provide feedback and interpretive information to our students about their assessment performances. We were gratified to offer this to all spring 2009 sophomores and fall 2008 entering first-year students who had completed the QR and SR. The email informed the students that they could log on to their JMU account on eCampus to view their QR and SR test scores. This portion of eCampus is titled “Test Score” and lists several varieties of student test scores such as: SAT, ACT, AP, Math Placement, Foreign Language Placement, and now their QR and SR scores. The students were instructed to write their QR and SR scores down and then click on the web link provided at the bottom of the test score summary page. The web link directed students to a
CARS webpage that allowed them to select from two kinds of interpretive information: normative (NR), criterion-referenced (CR) feedback, or both. Contact information was provided at the bottom of the email for additional help; there were many phone call and email exchanges.

While we were not able to monitor how many and which students accessed their eCampus accounts, we were able to track how many students went to the CARS website on each day as well as which type of interpretive information was selected. As described above, the following four cohorts were invited to view their scores and gain interpretive information: 1) the experimental feedback condition; 2) the control condition; 3) the remaining students who completed the NW-9 in the spring of 2009 who were involved in a separate study; and 4) all entering first-year students who completed the NW-9 in the fall of 2008. It should be noted, that following the delivery of the email to the second group of students, the order in which the links were presented on the webpage was reversed in order to examine whether the presentation order was affecting the type of feedback students selected. The number of unique visitors to the feedback landing page and to each of the feedback pages was tracked and recorded daily to determine if there were differences between groups in the number of students seeking feedback and within groups in the number of students choosing to seek normative or criterion-referenced feedback.

Results and Interpretation

The first stage of analysis compared the experimental and control group scores to determine if there were significant differences on the Importance or Effort subscales of the SOS. No significant differences were observed for examinee motivation. Table 1 presents the descriptive statistics and t-test results for these analyses. It should be noted that the SOS was completed at the end of the assessment testing session. Scores can range from 5 to 25.
Table 1

*SOS Subscore Descriptive and t-test Results*

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Mean</th>
<th>SD</th>
<th>t</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Importance Score</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Experimental</td>
<td>215</td>
<td>14.47</td>
<td>4.09</td>
<td>-0.919</td>
<td>.358</td>
</tr>
<tr>
<td>Control Group</td>
<td>314</td>
<td>14.80</td>
<td>4.06</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Mean</th>
<th>SD</th>
<th>t</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Effort Score</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Experimental</td>
<td>216</td>
<td>18.11</td>
<td>3.60</td>
<td>-1.618</td>
<td>.106</td>
</tr>
<tr>
<td>Control Group</td>
<td>316</td>
<td>18.59</td>
<td>3.12</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Similar analyses were conducted on the total Scientific Reasoning test and the Quantitative Reasoning test scores. These results are presented in Table 2. Students in the experimental condition scored significantly higher than students in the control condition on both the SR and the QR tests. However, while the differences vary from what would reasonably be expected by chance, the effect sizes were small for both comparisons ($d = .21$ for both analyses). The practical significance of the differences can be reasonably debated; however, it should be noted that the experimental intervention was remarkably subtle.

Table 2

*Descriptive Statistics and t-test Results for Scientific and Quantitative Reasoning Tests*

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Mean</th>
<th>SD</th>
<th>t</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Scientific Reasoning Test</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Experimental</td>
<td>218</td>
<td>35.88</td>
<td>6.54</td>
<td>2.35</td>
<td>&lt;.05</td>
</tr>
<tr>
<td>Control Group</td>
<td>316</td>
<td>34.60</td>
<td>5.95</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Mean</th>
<th>SD</th>
<th>t</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Quantitative Reasoning Test</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Experimental</td>
<td>218</td>
<td>18.03</td>
<td>3.60</td>
<td>2.40</td>
<td>&lt;.05</td>
</tr>
<tr>
<td>Control Group</td>
<td>316</td>
<td>17.16</td>
<td>4.14</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
An additional caveat is necessary at this point. While reviewing the room set up for the experimental and control rooms, it became apparent that the rooms assigned to the control group were all computer labs. Therefore, the ‘significant’ results observed may be attributed to a difference in testing modality.

We have occasionally observed differences between paper and pencil and computer-based performances, and these differences have usually advantaged the paper and pencil format. For first year Fall 2007 students, the effect size for the total test was .05; for Spring 2008 sophomores it was $d = -.03$, with the computer-based students performing slightly better than the paper and pencil group. Results for the QR test yielded the following effect sizes: first year students in Fall 2007 ($d = .28$); sophomores in Spring 2008 ($d = .02$). More research is necessary to more fully interpret these results. However, it would be imprudent to suggest that the results from this study did not include this confound.

The realities of competing for limited space and time on busy Assessment Days that must capture valuable data for five general education clusters, much of Student Affairs, and doctoral dissertations were evidenced here. The decision was made to relegate the control group to the labs, because this was a ‘one-time’ study, and it would be too expensive to modify the coding for the computer-based testing administration. Clearly, ideal study design was thwarted [even for the Executive Director of the Center!].

Table 3 displays the daily number of unique student visitors to each of the three web pages for each of the four groups after receiving the feedback availability email. The pattern of website visits suggests quite a bit of activity on the day of email delivery with traffic slowing over the weekend. A slight increase in activity is observed following the weekend.
Figure 1 presents the total number and percentage of students who visited each of the three web pages in the week (7 days) following email delivery. It is noteworthy that after the links to the NR and CR referenced pages were reversed at the Feedback Landing Page, the preference for the NR information disappeared. The CR interpretive information was now more highly sought, though clearly not to the extent of the previous groups’ preference for the NR information. The data suggest that an order effect was present. It was hypothesized that students were selecting the first link they saw, the NR link. For the first two groups, approximately three or four times as many students chose to seek normative feedback as opposed to criterion-referenced feedback. Following the reversal of link presentation on the FLP, more of the students in the third and fourth email groups selected CR interpretive information. However the preference was much less pronounced.
Figure 1. 
*Percentage of student visitors to each of 3 web pages during the week following email delivery, by group.*

![Graph showing visits to web pages by group](image)

Figure 2 provides visual displays of the numbers and percentage of visits to each of the three web pages for each of the four student groups. The study experimental group visited the web pages quite a bit more than the other groups. This may be due to the recency of the assessment activities compared to the other groups, though the time difference was only one week. The group that was not a part of this study, the ‘other’ group, was randomly assigned to participation in another motivation study. They seemed to be very interested in receiving Criterion Referenced interpretive information. Figure 2 makes it clear that the extent of their interest far exceeds any other group. The order effect for feedback request clearly influenced this preference, but the magnitude of the preference is quite overwhelming. It may be that the behavior observed for this group is a function of the intervention they received.
Figure 2. **Total number of unique student visitors and proportion of four groups by webpage**

**Feedback Landing Page**

- Experimental: 134
- Control: 122
- Other (Sp. 2009): 269
- Fall 2008: 485

**Norm-Reference Feedback Page**

- Experimental: 98
- Control: 106
- Other (Sp. 2009): 159
- Fall 2008: 261

**Criterion-Reference Feedback Page**

- Experimental: 27
- Control: 32
- Other (Sp. 2009): 171
- Fall 2008: 311
This study provided some rather interesting results regarding examinee motivation and performance in low-stakes testing conditions as well as new information about student preferences for types of test feedback. We need to know a great deal more about whether or not offering feedback provides any incentive that might impact student motivation or performance in low stakes conditions. It would seem logical that offering feedback cannot hamper motivation or performances.

Upon reflection, it not surprising that SOS scores were not impacted by the subtle promise of feedback. The SOS was administered at the end of an over 2 hour testing session. Perhaps if the SOS were administered immediately following the NW-9, we might have seen significant and practical differences. Other papers in this symposium employed this methodology.

It should also be noted that concerted research on examinee motivation has been conducted at this campus for well over 10 years, and we have witnessed substantive improvements in student participation and motivation due to our many interventions. How much more can we hope to observe, particularly with such a modest ‘treatment’? It may be that at another institution, where little has been done to improve student motivation, such a meager treatment might have more of an impact.

Something must be said about the ‘other’ sophomore group. These students were randomly assigned to rooms allocated for a separate motivation study. It may be that their treatment had a greater impact on their desire to learn more about their scores. This ‘other’ group clearly was not a control group. Their participation rate in going to the Feedback Landing Page was stronger than the Control group, despite an additional week delay in receiving their invitation to learn about their test scores and feedback. They also sought feedback at a greater rate than the first-year Fall 2008 student group. While their requests for NR interpretation was lower than the
control group, this was undoubtedly due to the order of presentation of the link on the feedback landing page. This conclusion appears warranted when the ‘other’ group’s requests for the CR information was overwhelmingly more numerous than any other group. The Fall 2008 entering students also requested the CR information in greater numbers than those requesting NR information. We will continue to study more carefully the impact of the offer for feedback to students in low-stakes testing conditions. The addition of type of feedback is of real interest to us. Our previous focus groups on feedback suggested that students would prefer to learn about their scores in relation to other students. One student indicated that learning about their scores in relation to faculty expectations might be disappointing if they did not achieve them.

We are currently preparing to send the emails to our Fall 2009 entering first-year students so that they can learn about their QR and SR scores and select interpretive information. These students had a slightly different set of instructions. Not only were they informed that they would be able to learn their scores and receive two types of interpretive information, they were also told that when they became eligible for Assessment Day again, we would make every effort to assure that they took the same exams again so they could see how much they had grown and developed over time. The overall results for the entering first year Fall 2009 class are really quite stunning. They are scoring like sophomores! Stay tuned!

References


Wise, S.L. (in press). Strategies for managing the problem of unmotivated examinees in low-stakes testing programs. Accepted for publication in Journal of General Education.
Appendix A: Test Instructions for Experimental Group

Welcome to Assessment Day!

This test is designed to assess your quantitative and scientific reasoning. At JMU we define these as thinking processes for obtaining and evaluating knowledge of the natural world. You will have 60 minutes to complete the 66 multiple-choice items on this test. You will have a piece of scrap paper to help you during this test. If you need more scrap paper, raise your hand. Make sure to read all test directions carefully and answer the items to the best of your ability.

For the first time, we are now able to provide you with feedback on your performance on this test using eCampus. You will receive an email in March that will inform you that your scores will be available in eCampus. Using your JMU ID, you can log onto eCampus and be able to review your scores and learn more about how to interpret your scores. You will be able to find out how you scored in relation to other students and how you scored in relation to content mastery. JMU is pleased to make this information available to you in such a convenient way.

Thank you in advance for your effort and attention to this important test.
Dear 

Thank you for your participation in JMU’s Spring Assessment Day. As you may recall, you completed JMU’s Quantitative (QR) and Scientific Reasoning (SR) instruments on that day. You were also informed that you would have the opportunity to receive feedback on your scores after the tests were scored.

I am pleased to inform you that the QR and SR tests have been analyzed, and your scores are now available to you on eCampus. When you log onto eCampus, you can now see your QR and SR test scores. If you are interested in learning more about those scores, write them down and click on the link at the bottom of the table. This link will take you to the Center for Assessment and Research Studies (CARS) test information webpage. You will find two links, one will link to information about how your score relates to other students' scores, and the other will link to information about how your score aligns with JMU faculty's expectations. Your scores and interpretive information will also be available to your advisor, should you wish to discuss them.

Thank you again for your participation in Assessment Day. JMU uses this information to make decisions about how to continue to improve our educational programs.

Best wishes,
Donna L. Sundre
Professor of Graduate Psychology/
Executive Director,
Center for Assessment and Research Studies (CARS)
821 S. Main St. [MSC 6806]
James Madison University
Harrisonburg, VA 22807
Phone: 540 568-3483
FAX: 540 568-7878
www.jmu.edu/assessment/
Appendix C: Email sent to other students informing them of test score availability

Dear 

Thank you for your participation in JMU’s Spring Assessment Day. As you may recall, as part of your Assessment Day activities, you completed JMU’s Quantitative (QR) and Scientific Reasoning (SR) instruments on that day.

I am pleased to inform you that the QR and SR tests have been analyzed, and your scores are now available to you on eCampus. When you log onto eCampus, you can now see your QR and SR test scores. If you are interested in learning more about those scores, write them down and click on the link at the bottom of the table. This link will take you to the Center for Assessment and Research Studies (CARS) test information webpage. You will find two links, one will link to information about how your score relates to other students' scores, and the other will link to information about how your score aligns with JMU faculty's expectations. Your scores and interpretive information will also be available to your advisor, should you wish to discuss them.

Thank you again for your participation in Assessment Day. JMU uses this information to make decisions about how to continue to improve our educational programs.

Best wishes,
Donna L. Sundre
Professor of Graduate Psychology/
Executive Director,
Center for Assessment and Research Studies (CARS)
821 S. Main St. [MSC 6806]
James Madison University
Harrisonburg, VA 22807
Phone: 540 568-3483
FAX: 540 568-7878
www.jmu.edu/assessment/