How reliable are our assessment data?:

A comparison of the reliability of data produced in graded and un-graded conditions.

Anthony R. Napoli & Lanette A. Raymond

Office of Institutional Research and Assessment

Suffolk County Community College

Send correspondence to:

Anthony R. Napoli, Ph.D.
Office of Institutional Research & Assessment
Suffolk County Community College
533 College Road, Selden, NY 11784

email: napolia@sunysuffolk.edu
voice/audix: 631-451-4842

Acknowledgement

This is a revised version of a paper presented at the Annual Forum of the Association for Institutional Research, Tampa, 2003; and was the recipient of the National Council of Research and Planning's Award for Best Paper. This research was supported by a grant from the Institute for Community College Development at Cornell University. The authors wish to thank Ms. Kathleen Massimo, Office of Institutional Research and Assessment at Suffolk County Community College, for her assistance in preparing this manuscript.

Abstract

Motivating students to perform well on assessment tests is difficult when students know the results have no academic consequence. The present study evaluates the influence of assessment context (graded v non-graded) on the reliability of an assessment measure. Results indicate the graded condition produces higher reliability (r = .71) than the non-graded condition (r = .29), which leads to unacceptably low reliability. Moreover, the graded condition produces significantly higher scores (M = 64%), than the non-graded condition (M = 43%). Only students in the graded condition (41%) obtained passing scores of 70% or above.

Key Words: assessment, higher education, reliability, authentic assessment, embedded assessment, student motivation

In American higher education, outcomes assessment is a process of examining institutional effectiveness for accountability to a variety of stake holders (e.g., regional accrediting agencies, professional accrediting bodies, and state education departments). One of the problems faced by institutions implementing outcomes assessment for the purposes of course review, program review, general education evaluation, or accreditation - rather than for awarding student grades - is motivating students to take the test seriously and to perform their best. This is particularly difficult when students know that the results of the assessment have little or no bearing upon them. The literature (Duvall, 1994; Warren, 1988) suggests that subjecting students to assessment tests that have no personal meaning and require giving up time and effort is likely to lead to resentment and less than maximum-performance efforts on the measurement.

Several motivational strategies that do not link performance on assessment measures to academic outcomes (e.g., course grades or graduation) have been employed to encourage students to perform their best (Nichols, 1995a; 1995b; Duvall, 1994). These strategies typically involve appealing to students to assist the institution in improving its curriculum. More creative alternatives have included practical or tangible rewards such as priority registration, preferred parking privileges, free lunch, gift certificates to the college bookstore, or college apparel. Unfortunately, there is no reliable evidence that these strategies do indeed motivate maximal student performance. As noted by Nichols (1995a, p. 47), “An appeal to the good nature of student body has, in most instances, fallen on deaf ears and few students take the assessment seriously once they realize that there is nothing in it for them.”

Based on a series of case studies, Nichols (1995a, 1995b) proposes that the most effective approach to motivating students is to integrate assessment activities into existing academic procedures; for example, embedding assessment into class work, a capstone course, or other

required academic endeavors such as classroom exams. Embedding assessment within a class motivates students to do their best "for the sake of their grade." Performance differences between graded and non-graded approaches on assessment measures have not been systematically explored. Moreover, the effects of these approaches on the reliability and validity of assessment data must also be examined.

The present study was conducted in response to state-mandated[1] General Education outcomes assessment; an issue imminently facing many institutions. During the development of appropriate assessment measures it became apparent that in order to report reliable, valid, and representative results to the state, the institution would need to invest in strengthening assessment methodology – both in terms of selected or designed measurement instruments and in terms of administration or implementation procedures. Although educational researchers have theorized to some degree on the impact of implementation procedures on outcomes, a dearth of published data on the subject led us to conduct the present study. The focus of the study is to evaluate the influence of assessment context (graded v non-graded) on the reliability and validity of assessment data. The graded condition should motivate consistent optimal performance whereas, the non-graded condition lacks such motivation and performance should be less consistent and suboptimal. Therefore, it is hypothesized that data collected from the graded condition would be both reliable and valid, but data collected from the non-graded condition would be unreliable, thereby precluding validity.

[1] State University of New York Assessment Initiatives. See http://www.cortland.edu/gear/SUNYassmt_initiative.pdf

Methods

Sample

Two groups of community college introductory psychology students were administered a multiple-choice assessment exam under different contextual conditions - a graded condition (n = 46) and a non-graded condition (n = 34). These groups did not differ significantly on any of the commonly measured demographics. They were similar in age and gender, as well as incoming academic skills, reflected by their high school averages, overall college averages, and scores on the College Placement Test's Reading Comprehension test (CPT-R; College Entrance Examination Board, 1990) collected when students were admitted to the college. See Table 1. Having ruled out subject-specific differences to which group differences could be attributed, group comparisons can be interpreted with greater confidence.

Table 1. Sample demographics by group.

	Ungraded Group			Graded Group				
	Mean	SD	N	Mean	SD	N	df	t
High School Average	78.7	6.5	34	77.2	6.1	46	78	1.04
Overall College GPA	2.7	1.1	34	2.6	1.0	46	78	0.17
CPT-R	81.1	17.9	25	79.5	19.3	36	59	0.32
Age	21.4	6.6	34	23.1	7.2	46	78	-1.07
Gender (percent female)	63.1	-	34	62.5	-	46	Z =	0.54

Procedure

The participating sections of Introductory Psychology are essentially equivalent. Introductory Psychology is a reading-intensive lecture course. Tests are based largely on the textbook. Course syllabi are scrutinized by the department chair to ensure a uniform set of core learning goals and objectives are taught in all Introductory Psychology courses within the department. Textbooks are selected on this basis as well. These core learning goals and objectives are represented in the content of the present study's measure (see *Measures*, below). Students in both conditions were exposed to the same set of course topics during the semester.

The two assessment conditions were designed to represent 1) the most common assessment context, which provides little motivation usually in a uniform motivational "speech" and assumes general preparation throughout the course or program rather than specific test preparation, in the non-graded condition, and 2) the more ideally motivated assessment context in the graded condition, (as presented in the discussion above) allowing for external validity of the results. In the graded condition assessment items were embedded into a cumulative end-of-term final exam -- 20 assessment items within a 100 item final. In the non-graded condition the 20 assessment items were given as a stand-alone set at the end of the term, one in which no final exam was scheduled. Students in the non-graded condition were given a brief motivational "speech"[2] encouraging them to answer the assessment items to the best of their ability in order to assist the college in improving its academic programs; similar to motivational introductions

[2] As part of our commitment to quality education, the faculty of _____ County Community College is conducting a series of outcome assessment studies to better understand the effectiveness of our academic programs. The studies will include the assessment of student's basic knowledge in a variety of core disciplines including psychology. To carry out the assessment, members of the psychology faculty have developed a short multiple-choice test of basic facts and concepts covered in Introduction to Psychology. Please assist us in this important work by answering, to the best of your ability, the following questions.

typical in voluntary proctored assessment contexts. No specific instructions were provided regarding guessing; though it is unlikely students refrained from guessing as there were no missing data in the either condition. In the non-graded condition the students were informed that their performance on the exam would have no influence on their course grade (a fact reinforced by the anonymity of the forms and voluntary participation). Students in both conditions were allowed to leave when they completed the test.

Measures

The primary measure was a 20-item multiple-choice assessment instrument (PsyOA) developed for use in introductory psychology classes for the purpose of General Education assessment by a committee of psychology faculty. This committee formally defined the course objectives and developed a balanced set of items to represent the breadth of the course and specifically address each of these objectives. Each item included 4 answer options. Course grade-point-averages (GPA) and midterm- and final-exam scores were obtained for students in the graded condition to examine the concurrent validity of the assessment measure. Because introductory psychology is a reading intensive course, scores on the CPT-R for students in this group were also collected and used to examine the construct validity of the measure.

Results

Reliability

The internal consistency reliability of the PsyOA measure was independently evaluated for each group using the Kuder-Richardson 20 formula (KR-20; Rosnow & Rosenthal, 2002). The results for the non-graded group indicate that under this condition the measure has unacceptably low internal consistency reliability (r_{tt} = .29). Conversely, the reliability for the graded group was substantially higher and within an acceptable range (Mehrens & Lehmann, 1973; Nunnally & Bernstein, 1994) for a new assessment measure (r_{tt} = .71).

Between group performance differences

A comparison of the group means for graded and non-graded students on the PsyOA measure shows that the graded condition produces significantly higher scores ($t(78) = 5.62$, $p < .001$). Specifically, the graded students had an average PsyOA score of 12.48 correct items ($SD = 3.48$) or 62.4%, whereas the non-graded students scored an average of 8.59 items correct ($SD = 2.36$) or 42.9%. Not only is this difference between the two groups statistically significant, the effect size ($d = 1.27$), based on Cohen's criteria (Cohen, 1992), is large. Alternately stated, embedding assessment questions in a graded exam motivates substantially superior performance when compared to a non-graded situation. An examination of the proportion of students that obtained grades of 70% or above (a "C" grade or above) on the PsyOA -- arguably the point of minimum competency -- was also conducted. The results indicate that none of the non-graded students obtained a score of 70 or above, while a significantly higher 41.3% of the graded students obtained scores of 70 or above ($Z = 5.69$, $p < .001$).

Validity

To assess the criterion-related concurrent validity of the PsyOA measure among the graded students, a correlational analysis was conducted on the assessment test scores and students' term averages and course GPAs. The results show that performance on the PsyOA test was strongly and significantly related to both term average (r = .849, p < .001) and course GPA (r = .792, p < .001). It is interesting to note that assessment scores in the graded group were highly correlated with scores on the remaining 80 items in the final (r_{tt} = .72; M_{80} = 70%, SD_{80} = 14.9%).

To establish the construct validity (in terms of convergent validity) of the assessment test scores, student's PsyOA scores, GPA and term averages were correlated with their CPT-R reading comprehension test scores (see Table 2). The results indicate that the CPT-R is significantly related to the PsyOA (r = .401, p < .01). CPT-R scores were also significantly related to GPA (r = .450, p < .05) and term averages (r = .476, p < .01).

Table 2. Correlation between assessment measure, course outcomes, and CPT-R scores.

	(2)	(3)	(4)	Mean	SD	n
Assessment grade (1)	0.849 ***	0.792 ***	0.448 *	62.39	17.41	46
Term Average (2)		0.954 ***	0.476 **	81.41	12.12	46
Course GPA (3)			0.450 *	2.88	0.91	46
CPT-R score (4)				80.63	18.04	30

* p < .05, ** p < .01, p < .0001

Discussion

Summary of reliability and performance data

Based on the results reported above, it is reasonable to conclude that when student performance on assessment measures is not linked to course outcomes (i.e., course GPA or pass/fail outcomes), due to a lack of motivation, their scores cannot serve as reliable indicators of their true learning or mastery of the curriculum. However, when scores on assessment measures are linked to course outcomes, students will be motivated to maximally perform and their scores can serve as reliable indicators of learning or mastery of the curriculum.

Summary of validity studies.

The PsyOA measure was designed to assess students' knowledge of the basic facts and concepts presented in introductory psychology. The concurrent validity of the measure was assessed by correlating scores on the PsyOA with students' term averages based on two in-class objective examinations and overall course GPA. The results indicate that performance on the assessment is indeed strongly associated with knowledge of the facts and concepts and overall performance in the class. Despite having acceptable reliability, if the results of the concurrent validity analyses had failed to demonstrate a strong linkage between the assessment measure and other independent indicators of academic achievement within the course, the meaningfulness and utility of the measure would be questionable.

Findings for the convergent validity further strengthen our confidence in the measure. Because introductory psychology is a reading intensive course it was believed that students who possess higher levels of reading comprehension abilities would achieve greater success in the course as indicated by their PsyOA performance. Previously, Napoli and Wortman (1995) had

observed that reading comprehension skills were significantly related to academic success (i.e., GPA) in introductory psychology. Consistent with this finding, as well as the hypothesized relationship between reading comprehension and PsyOA, the results obtained in the present study demonstrate that PsyOA is indeed significantly related to students' reading comprehension skills. This is important as it informs those who use the PsyOA with successive cohorts that later cohort comparisons of performance on the measure should be made controlling for differences in students basic reading comprehension skills.

Conclusion

Consistent with the findings presented above, Aronson, Brewer, and Carlsmith (1985) point out that even field studies can be contrived and artificial, with little real-world relevance. They refer to the similarity of research events to real-world occurrences as "mundane realism." Cozby (2001) notes that studies low in mundane realism, that bear little similarity to real-world events, or tasks that have no impact on the participants and which fail to engage the interest or involvement of the participants are not likely to yield valuable results. Such unrealistic exercises may also produce a degree of resentment among student participants, which may exacerbate performance deficiencies as well as impact the reliability of results.

In the present study, the non-graded assessment condition failed to produce sufficient mundane realism. Consequently, the performance of these students is (predictably) unreliable and serves as a poor indicator of their actual or learning or mastery of the curriculum. Assessment data obtained under this condition underestimate the true knowledge and ability of students, which leads to erroneous conclusions concerning student learning. The graded condition produced high mundane realism, and motivated students to perform to the best of their

ability. Data obtained under this condition can be reliable and the results, if derived through appropriate sampling techniques and based on an adequate number of students, are likely to generalize well to the larger population and provide valid information concerning the actual learning or mastery of the curriculum.

Colleges that have successfully involved students in testing have established assessment testing as an integral part of the curriculum (Duvall, 1994; Warren, 1988). Students do not question the use of course tests, quizzes or exams as a part of college work. Inconsequential assessment testing, conducted separate from the teaching-learning-grading process, is an unrealistic quixotic exercise, unlikely to elicit a maximum-effort response from the students. Consequently, test results obtained under such circumstances are poor indicators of the students' true knowledge and ability. If unmotivated assessment is acknowledged for producing unreliable data, the inconsequential mass-testing paradigm colleges rely upon nationwide will require review.

References

Anastasi, A. & Urbina, S. (1997). *Psychological testing (7th ed.).* Upper Saddle River, NJ: Prentice Hall.

Aronson, E., Brewer, M., & Carlsmith J.M. (1985). Experimentation in social psychology. In G. Lindzey & E. Aronson (Eds.), *Handbook of social psychology (3rd ed.).* New York: Random House.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112,* 155-159.

College Entrance Examination Board and Educational Testing Service. (1990). *Coordinator's Guide for Computerized Placement Tests, Version 3.0.* Princeton, NJ.

Cozby, P.C. (2001). *Methods in behavioral research (7th ed).* Mountain View, CA: Mayfield Publishing.

Duvall, B. (1994). Obtaining student cooperation for assessment. *New Directions for Community Colleges, 88(winter),* 47-52.

Gay, L. R. (1996). *Educational research (5th ed).* Upper Saddle River, NJ: Prentice-Hall.

Mehrens, W. A. & Lehmann, I. J. (1973). *Measurement and evaluation in education and psychology.* New York: Holt, Rinehart and Winston, Inc.

Napoli, A.R., and Wortman, P. (1995). Validating college placement tests. *Journal of Applied Research in the Community College, 3(2),* 143-151.

Nichols, J. (1995a). *Assessment case studies: Common issues in implementation with various campus approaches to resolution.* Bronx, NY: Agathon Press.

Nichols, J. (1995b). *A practitioner's handbook for institutional effectiveness and student outcomes assessment implementation* (3rd ed.). Edison, NJ: Agathon Press.

Nunnally, J. C. & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York:

McGraw-Hill.

Pittenger, D. J. (2003). *Behavioral research: Design and analysis.* New York: McGraw-Hill.

Rosnow, R.L., and Rosenthal, R. (2002). *Beginning behavioral research: A conceptual primer* (4th ed.). Upper Saddle River, NJ: Prentice Hall.

Warren, J. (1998). Cognitive measures in assessing learning. *New Directions for Institutional Research [No. 59, Implementing Outcomes Assessment: Promise and Perils], 1(3),* 29-39.