Use of the Purdue Cafeteria System for Instructor and Course Evaluation

A Report by the Committee to Evaluate Teaching Assessment Instruments
Department of Psychology
Eastern Illinois University
26 January 2011

Committee Members:
Gary L. Canivez, Ph.D., Chair
Wesley D. Allan, Ph.D.
Kristin N. Johnson, Ph.D.

Use of the Purdue Cafeteria System for Instructor and Course Evaluation

Each semester, Eastern Illinois University instructors are required to administer student course evaluations for each of their courses. At the present time, the Department of Psychology, like many departments at EIU, uses the Purdue Cafeteria System, which includes 194 possible item stems for which students might rate potential characteristics of the instructor, course, or activities. Items are rated on a five-point scale ranging from Strongly Disagree to Strongly Agree with "Undecided" at the midpoint. The 194 items are nested within the following 19 areas:

1. Clarity and Effectiveness of Presentations
2. Student Interest/Involvement in Learning
3. Broadening Student Outlook
4. Teaching/Learning of Relationships and Concepts
5. Instructor Provides Help as Needed
6. Providing Feedback to Students
7. Adapting to Individual Differences
8. Respect and Rapport
9. Course Goals or Objectives
10. Usefulness/Relevance of Content
11. Discussion
12. Exams and Grades
13. Assignments
14. Media: Films
15. Team Teaching
16. General Method
17. Laboratory
18. General Student Perceptions
19. Miscellaneous Items

Eastern Illinois University adopted five of the 194 Purdue items to represent a core set of questions that are administered to all students in each class. These are listed below.

University Core Questions
Instructor demonstrates command of the subject/discipline.
Instructor effectively organizes material for teaching/learning.
Instructor is readily available outside of class.
Instructor presents knowledge or material effectively.
Instructor encourages and interests students in learning.

Each department may select a set of questions for use within the department and the Department of Psychology has selected the five questions below to be included in all student course evaluations.

Departmental Core Questions
My instructor seems well prepared for class. (DC)
My instructor has stimulated my thinking. (DC)
When I have a question or comment I know it will be respected. (DC)
This course has clearly stated objectives. (DC)
I would recommend this instructor to a friend. (DC)

Individual faculty may also select any number of additional items from among the remaining 184 Purdue items to include in the course evaluation. This allows a faculty member to gather information he or she feel may be helpful in assessing his or her instruction and/or course.

The course evaluation may be administered in class (proctored by a student with instructor absent from the classroom) or may be administered online whereby each student in the course is sent an e-mail with a URL link to the secure web based form. Systematic investigation of differences in ratings between the in-class and internet based ratings has not yet been completed.

PSYCHOMETRIC FOUNDATIONS FOR INSTRUMENT (TEST) USE

In psychology and education (and other scientific disciplines), instruments used to gather information in research and applied settings *must* be evaluated in terms of basic psychometric properties and there are a myriad of methods and procedures for such investigations. Each psychometric property and method answers a different question and provides information for use of the instrument.

Two critical documents are fundamental for the use of instruments in psychology in both research and applied settings: *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA, APA, NCME], 1999) and *Ethical Principles of Psychologists and Code of Conduct* (American Psychological Association [APA], 2010). The *Standards* provide numerous guidelines for considering reliability and validity of test scores that should be applied to test scores. Such guidelines apply to test authors and publishers, but ultimately it is the test *user* who must decide which test scores, comparisons, and procedures possess sufficient evidence of reliability and validity to report and interpret. Test scores that do not possess adequate reliability, validity, and utility (individual application) will lead the test user to make inaccurate and inappropriate inferences about the individual when interpreting those test scores and comparisons. Such inaccurate and inappropriate inferences may well lead to recommendations that are erroneous. Weiner (1989) cogently noted in order to practice in an ethical manner, psychologists must "(a) know what their tests can do and (b) act accordingly" (p. 829). Numerous ethical standards also concern the use of tests and measurement procedures (APA, 2010).

Reliability of Scores

In considering the three fundamental psychometric properties, all tests and measures must first demonstrate acceptable levels of reliability. Broadly defined, test reliability refers to the consistency or precision of measurement. There are several methods to examine reliability and answer different questions. These include *internal consistency (item homogeneity), test-retest (stability), alternate forms (equivalence),* and *interrater agreement (consistency)*. If scores from a measure lack sufficient levels of reliability, it *cannot* be valid because it will possess too much error variance and lack sufficient precision. Reliability is a necessary, but not sufficient condition for test use. Numerous psychometric experts have indicated that reliability coefficients for tests or measures must meet or exceed a criterion of .90 in order to be used for individual interpretation and decision–making (Aiken, 2000; Guilford & Fruchter, 1978; Nunnally & Bernstein, 1994; Ponterotto & Ruckdeschel, 2007; Salvia & Ysseldyke, 1988, 2001). For group decisions or research purposes the reliability coefficient may be as low as .70, but use of instruments with lower reliability estimates comes with a cost of significant reduction in power. It must be stated that reliability *is not* a property of the test or measure but relates to the *scores* obtained on the test, that are yielded with a particular sample, in a particular setting, and at a particular time. Thus, it is expected that a range of reliability coefficients will be obtained through psychometric examinations, no single one being the sole determinant.

The *Standards* includes numerous standards pertaining to reliability and the ones that seem most germane to the use of the Purdue Cafeteria System include the following:

Standard 2.1: For each total score, subscore, or combination of scores that is to be interpreted, estimates of relevant reliabilities and standard errors of measurement or test information functions should be reported.
Standard 2.5: A reliability coefficient or standard error of measurement based on one approach should not be interpreted as interchangeable with another derived by a different technique unless their implicit definitions of measurement error are equivalent.

Standard 2.14: Conditional standard errors of measurement should be reported at several score levels if constancy cannot be assumed. Where cut scores are specified for selection or classification, the standard errors of measurement should be reported in the vicinity of each cut score.

The application of reliability estimates for score interpretation relates to a score as an *estimate* of the person's *true score* on that measure and interpretation of scores must include a confidence interval to illustrate the error in measurement present in *all* test scores. This helps guard against the *illusion* of a score as being *the* score of the person. The confidence interval (obtained score CI *or* estimated true score CI) will be constructed based on the standard error of measurement yielded from one or more of the reliability estimate methods. Confidence interval estimates based on the internal consistency estimate for a measure will provide the best–case scenario for precision in measurement but precision may actually be worse depending on the method, requiring a larger confidence interval.

In the case of the Purdue Cafeteria System, it appears that estimates of internal consistency and interrater agreement are likely the two most important methods for providing estimates of reliability of scores. One could also see a role for estimates of short–term test–retest stability.

Validity of Scores

While reliability is a necessary condition for a test or measure, validity research provides information on *how* scores or test results are to be interpreted. The meaning attributed to a test score, or the inference made about someone based on the test score, requires strong evidence of validity. Like reliability, validity is not a unitary concept and includes many different methods each designed to provide evidence for certain interpretations. While the Trinitarian model of validity (Content Validity, Criterion–Related Validity, Construct Validity) prevailed for over 50 years; more recently Messick (1989, 1995) proposed a model that is more contemporary and his six distinguishable components of construct validity are content, substantive, structural, generalizability, external, and consequential aspects of construct validity and all are reflected in the *Standards.* Sources of validity evidence illustrated in the *Standards include* (1) evidence based on test content, (2) evidence based on response processes, (3) evidence based on internal structure, (4) evidence based on relations to other variables, and (5) evidence based on consequences of testing.

The *Standards* includes numerous standards pertaining to validity and the ones that seem most germane to the use of the Purdue Cafeteria System include the following:

Standard 1.1: A rationale should be presented for each recommended interpretation and use of test scores, together with a comprehensive summary of the evidence and theory bearing on the intended use or interpretation.
Standard 1.2: The test developer should set forth clearly how test scores are intended to be interpreted and used.
Standard 1.3: If validity for some common or likely interpretation has not been investigated, or if the interpretation is inconsistent with available evidence, that fact should be made clear and potential users should be cautioned about making unsupported interpretations.
Standard 1.4: If a test is used in a way that has not been validated, it is incumbent on the user to justify the new use, collecting new evidence if necessary.

Utility of Scores

Utility, or more precisely diagnostic utility or efficiency, is dependent on a measure having adequate estimates of validity, particularly based on evidence from relations to other variables where distinctly different groups with respect to a construct produce significantly different scores on a measure of that construct. But while a test may show the ability to differentiate groups, application of test scores for making decisions about *individuals* requires an even stronger source of evidence. Distinct group differences are a necessary but not sufficient condition for the use of test scores for making *individual* decisions. This is because groups may have significantly different means but their distributions may overlap considerably and when applying a cut score to an individual there is, depending on the degree of overlap, some degree of false positive or false negative decisions. In the case of

individual decision making with respect to a test, there is a prediction for the individual based on a score (or scores) from a measure. That prediction may be a categorical judgment, for example "low, medium, or high," or a dichotomous judgment indicating presence or absence of some condition. It is also possible to make individual predictions with respect to a continuum but it is likely that some categorical decision will be made even with prediction on a continuum. Regardless, McFall (2005) correctly points out the necessity for assessing utility.

The *Standards* does not include standards pertaining specifically to diagnostic utility or efficiency as with reliability and validity but there are some areas that seem to relate to the use of the Purdue Cafeteria System. One area indicates "response–related sources of test bias" where "construct–irrelevant score components may arise because test items elicit varieties of responses other than those intended or can be solved in ways that were not intended" (AERA, APA, NCME, 1999, p. 78). Another area might be in that of fairness in selection and prediction and "when tests are used for selection and prediction, evidence of bias or lack of bias is generally sought in the relationships between test and criterion scores for the respective groups" (AERA, APA, NCME, 1999, p. 79). Investigation of variables that might be related to test bias should be conducted, particularly in high–stakes assessments.

ETHICAL PRACTICES IN TEST USE

It is sometimes said that in order to be a profession, that profession must have a code of ethics that guides the behaviors and practices of its members. Psychology is one such profession and the most recent amendments to the ethical code were published in 2010 (American Psychological Association [APA], 2010). Because psychologists are intimately involved in measurement, assessment, test development, and test use; there are a number of ethical standards that apply within this area.

Specific APA ethical principles that affect test score use and interpretations are listed below:

2.04 Bases for Scientific and Professional Judgments
Psychologists' work is based upon established scientific and professional knowledge of the discipline. (See also Standards 2.01e, Boundaries of Competence, and 10.01b, Informed Consent to Therapy.)

9.01 Bases for Assessments
(a) Psychologists base the opinions contained in their recommendations, reports, and diagnostic or evaluative statements, including forensic testimony, on information and techniques sufficient to substantiate their findings.

9.02 Use of Assessments
(a) Psychologists administer, adapt, score, interpret, or use assessment techniques, interviews, tests, or instruments in a manner and for purposes that are appropriate in light of the research on or evidence of the usefulness and proper application of the techniques.
(b) Psychologists use assessment instruments whose validity and reliability have been established for use with members of the population tested. When such validity or reliability has not been established, psychologists describe the strengths and limitations of test results and interpretation.

9.08 Obsolete Tests and Outdated Test Results
(a) Psychologists do not base their assessment or intervention decisions or recommendations on data or test results that are outdated for the current purpose.
(b) Psychologists do not base such decisions or recommendations on tests and measures that are obsolete and not useful for the current purpose.
(c) Psychologists retain responsibility for the appropriate application, interpretation, and use of assessment instruments, whether they score and interpret such tests themselves or use automated or other services.

PSYCHOMETRIC FOUNDATIONS AND ETHICAL STANDARDS SUMMARY

To summarize, all tests or measures *must* demonstrate empirical evidence for score reliability, validity, and utility in order that the scores are properly interpreted and used, where appropriate. Tests that do not have acceptable psychometric properties ought not be used or if used, used with extreme caution and qualifications presented when scores are reported and used. It is incumbent upon those promoting the use of specific tests, scores, or ratings for specific purposes to provide psychometric evidence for the test, scores, or ratings in question. Individuals, groups, and agencies are urged to use tests or measures appropriately. Appropriate use of tests begins with selection of measures based on *evidence* of acceptable score reliability for the specific use of the test (individual decision making vs. group/research purposes). For measures that possess acceptable score reliability, appropriate validity research provides the means by which one interprets the scores and derives meaning or makes inferences about the score or ratings. Such interpretations and inferences are specific to the research methods and settings and may not generalize to other settings. Finally, if the measure possesses strong validity, inspection of its *utility* for individual decision–making (high–stakes decisions such as selection, promotion, etc.) is necessary to be sure that a minimum of false positive and false negative decisions are made. It is an ethical responsibility of psychologists to adhere to these principles and help guard against inappropriate use of tests and measures.

PSYCHOMETRIC PROPERTIES OF THE PURDUE CAFETERIA SYSTEM

In an effort to obtain information regarding the procedure and criteria for the selection of the Purdue Cafeteria System for instructor and course evaluation at EIU, the committee solicited information from Karla J. Sanders, Ph.D., Director of the Center for Academic Support and Achievement at EIU. Her reply indicted that she did not have much information on the Purdue so she couldn't be of much help but noted that the Purdue was selected some time in the early 1980s but did not know the mechanism used for its selection. She stated, "I know that the instrument was validated by Purdue, but I don't have the information on that." Additionally, she noted that whether the Purdue or another department based survey was used, the core questions (noted above) must be used. Finally, in response to the committee's query regarding the collection of data for evaluation purposes by EIU, Dr. Sanders noted that no such aggregation was done.

A search of the EIU website provided no specific information on the Purdue with respect to why it was selected or evidence that its use at EIU had ever been examined for its psychometric fitness (reliability, validity, utility). Even within the Department of Psychology, there is no information as to any investigation as to the psychometric properties of the Purdue (reliability, validity, utility).

A search was conducted for peer reviewed research publications regarding the Purdue Cafeteria System in the psychology and educational literature in order to determine estimates of score reliability, validity, and utility that might help to judge the psychometric adequacy and guide Purdue use. Literature searches (PsychINFO, PsychARTICLES, ERIC, Academic Search Premier) for published studies regarding the Purdue Cafeteria System proved fruitless.

A Google search for the Purdue Cafeteria System provided several links to university web sites or reports. Among them was report by the Suffolk County Community College Office of Institutional Research (report date unknown) (http://instsrv.sunysuffolk.edu/strate.htm) that provided a detailed review of the literature on student ratings of instruction. It must have been produced sometime after 1998, as this is the publication year of most recent reference in the report. This report also noted the absence of published research regarding the Purdue Cafeteria System with respect to both reliability and validity.

The University of Southern Indiana posted on their web site a summary of their faculty senate committee report on student evaluation of faculty teaching (http://www.usi.edu/facsenate/StudentEvaluationofTeaching.asp) and it was noted that, "an evaluation of the Purdue cafeteria forms was conducted by James Divine in 1990. USI faculty was surveyed about the Purdue Cafeteria form in January 2002. These findings are presented in two documents: a summary of responses to the survey and written comments on the survey." No further information was available nor was any psychometric information provided regarding the Purdue.

Purdue University Calumet had a posting of a senate resolution (http://library.calumet.purdue.edu/Faculty_Senate/Documents/2004-2005/December/SD_04-04%20(Teaching%20Effectiveness).html) that noted in 2004, Purdue University Calumet was the only Purdue campus using the Purdue Cafeteria System and that it "is no longer a viable instrument for this campus." No information was available as to psychometric properties of the Purdue Cafeteria System.

Finally, the Google search produced a link for an article published in *Science* (Rodin & Rodin, 1972) that reviewed previous research regarding the Purdue Rating Scale (it is not known if this is identical to the Purdue Cafeteria System) and relationships to student grades and learning. The studies referenced by Rodin and Rodin (1972) that examined the Purdue Rating Scale were, at that time, dated (Elliot, 1950; Remmers, 1928, 1930; Remmers, Martin, & Elliot, 1949) and are especially so today and likely not particularly informative. Even the results of Rodin and Rodin (1972) may not apply to the present version of the Purdue Cafeteria System or specific items in use at EIU and in the Department of Psychology.

PSYCHOMETRIC PROPERTIES OF THE PURDUE CAFETERIA SYSTEM SUMMARY

Based on the dated information on the Purdue Rating Scale, which we do not know how it relates to the present Purdue Cafeteria System or items selected for use by EIU, and the lack of psychometric information regarding the Purdue Cafeteria System score reliability, validity, and utility, it is impossible to know to what extent information (scores) from it are sufficiently reliable, valid, or of utility for individual decision–making. The committee could find no empirical research in the published literature or in readily available reports via the internet reporting on the psychometric fitness of the Purdue Cafeteria System that would support its use. Also disconcerting is the fact that there is apparently no psychometric information regarding the Purdue Cafeteria System based on its use at EIU since the early 1980s or within the Department of Psychology since its adoption several years ago. With respect to the five core University items and the five Department of Psychology items there is no evidence for their specific score reliability, validity, or utility and given the few number of items it is unlikely that strong psychometric support would be obtained. In the absence of empirical data on reliability, validity, and utility; use of the Purdue Cafeteria System for providing an evaluation of a course or an instructor can only be based on belief, speculation, or conjecture.

Because there is a lack of information regarding the reliability, validity, or utility of the Purdue Cafeteria System scores (ratings), information from it *cannot* be meaningfully interpreted. Further, because of its formal use and statistical report, there is the *appearance* that it is measuring aspects of courses or instructors but interpretations are *illusory*. What do high ratings, medium ratings, or low ratings mean and what is the basis for determining whether a rating is high, medium, or low? Because there are no reliability data one cannot be sure that the obtained ratings are any better than chance. Because there are no validity data one cannot know what the Purdue ratings mean apart from the content or face validity of the items selected. Are Purdue ratings related to student outcomes (learning) and related to instructor and/or course quality? Are Purdue ratings influenced by student characteristics? Are Purdue ratings influenced by irrelevant faculty characteristics or by irrelevant student characteristics? In the absence of empirical evidence examining such issues it is impossible to know what or how such influences affect the Purdue Cafeteria System ratings. Most importantly, because there are no utility data, one cannot know the extent to which high ratings reflect effective instruction and/or a quality course or whether it reflects a false positive judgment based on construct–irrelevant variance. Likewise, one cannot know the extent to which low ratings reflect poor instruction and/or a poor course or whether it reflects a false negative judgment that might be based on construct–irrelevant variance. Is it possible for Purdue Cafeteria System ratings to be influenced by instructor behaviors in order to produce high ratings that are not reflective of actual instructional or course effectiveness? Without answers to these and other questions, inferences and interpretations of Purdue Cafeteria System ratings are subjective and will likely vary considerably between individuals considering them. Discussions with a variety of faculty at EIU have yielded different opinions of Purdue Cafeteria System ratings ranging from those who believe they reflect actual instructor effectiveness to those who believe that the content does not adequately measure effective elements of instruction. In the absence of evidence, any belief as to the utility of Purdue Cafeteria System ratings is merely that, belief

RECOMMENDATIONS

Based on the committee's search for empirical evidence in support of the Purdue Cafeteria System for instructor and course evaluation, and finding none, the following recommendations are offered to the faculty of the Department of Psychology and Chair of the Department of Psychology. Others who may be interested in the content of this report and its recommendations may include other university department faculty and department chairs, the Dean of the College of Sciences (and other Deans), the Provost and Vice President of Academic Affairs, and the President.

1. The Purdue Cafeteria System should be replaced with another instrument and procedure that has evidence for score reliability and validity in order that interpretations are correct, meaningful, and appropriate.

2. Selection of a replacement instrument to assess courses and instructors should be based upon a well thought out and articulated set of criteria for what aspects of instructional effectiveness and course quality should be assessed. Evaluation must begin with a definition for what qualities should be measured in order to select a measure that assesses the defined qualities. If no such measure exists then one must be created and investigated for score/rating reliability, validity, and utility before high–stakes decisions are made with scores from such an instrument.

3. Until adequate score reliability and validity is established within the Department of Psychology for *any* student rating of courses or instruction such ratings should be demoted within the DAC in order that they do not carry undue weight or influence in high–stakes decisions such as retention, tenure, promotion, and PAI. Further, the Department of Psychology Personnel Committee (DPC) and department Chair should include in formal evaluations of individual faculty statements of qualification and limitations regarding the reliability, validity, and utility of Purdue Cafeteria System ratings so that others above the department (Dean, UPC, Provost/VPAA) do not place undue influence or value on such ratings be they low, moderate, or high.

References

Aiken, L. R. (2000). *Psychological testing and assessment* (10th ed.). Needham Heights, MA: Allyn & Bacon.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

American Psychological Association. (2002, 2010 Amendments). *Ethical principles of psychologists and code of conduct.* Washington, DC: APA.

Cronbach, L. J. (1990). *Essentials of psychological testing* (5th ed.). Boston: Addison–Wesley.

Elliot, D. N. (1950). *Purdue Univ. Stud. Higher Educ., 70,* 5-.

Guilford, J. P., & Fruchter, B. (1978). *Fundamental statistics in psychology and education* (6th ed.). New York: McGraw–Hill.

McFall, R. M. (2005). Theory and utility—Key themes in evidence–based assessment: Comment on the special section. *Psychological Assessment, 17,* 312–323.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: Macmillan.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from person's responses and performances as scientific inquiry into score meaning. *American Psychologist, 50,* 741-749.

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

Ponterotto, J. G., & Ruckdeschel, D. E. (2007). An overview of coefficient alpha and a reliability matrix for estimating adequacy of internal consistency coefficients with psychological research measures. *Perceptual and Motor Skills, 105,* 997–1014.

Remmers, H. H. (1928). *School and Society, 28,* 59-.

Remmers, H. H. (1930). *Journal of Educational Research, 21,* 314-.

Remmers, H. H., Martin, F. D., & Elliot, D. N. (1949). *Purdue Univ. Stud. Higher Educ., 66,* 17-

Rodin, M, & Rodin, B. (1972). Student evaluations of teachers. *Science, 177,* 1164–1166.

Salvia, J., & Ysseldyke, J. E. (1988). Assessment in special and remedial education (4th ed.). Boston: Houghton Mifflin.
Salvia, J., & Ysseldyke, J. E. (2001). *Assessment* (8th ed.). Boston: Houghton Mifflin.
Weiner, I. B. (1989). On competence and ethicality in psychodiagnostic assessment. *Journal of Personality Assessment, 53,* 827–831.