NATIONAL FORUM OF TEACHER EDUCATION JOURNAL
VOLUME 19, NUMBER 3, 2009

Enhancing Validity of Critical Tasks Selected for College and University Program Portfolios

Pattie Johnston, PhD
Assistant Professor of Education
The University of Tampa
Tampa, FL

Karen Wilkinson, M.A.
Middle School Teacher for Social Sciences
Tampa, FL

ABSTRACT

University and college departments are often charged with creating assessment systems that measure student outcomes on identified objectives to meet accreditation standards. Some departments like Education have additional accreditation requirements because they have domain related accreditation agencies in addition to national or regional university agencies. Assessment systems usually consist of department objectives with assessments to measure student performance on each of the department goals. Students are often required to keep a portfolio of these assessment tasks. It is essential for assessment systems to have tasks that measure department objectives aligned directly with the stated objective. Departments have typically relied on faculty consensus to assure the desired alignment. Consensus can be difficult, an excessive amount of tasks may be identified or departments may want to spend a bit more effort to assure that their consensus is valid.

Introduction

Departments in accredited colleges and universities are usually required to assess student outcomes. This requirement is particularly true for departments with additional domain specific accreditation like Education. Departments are supposed to create assessments systems that identify student learning goals and the assessments given to measure acquisition of the stated goals. Data on student performance on the objective related assessments are typically tracked, aggregated and used to drive department improvement.

As mentioned above, it is of extreme importance for departments like Education to use student performance data to drive continual department improvement. National Council for Accreditation of Teacher Education **(**NCATE) 2000 standards require teacher education programs to assess pre-service teachers on departmental objectives *over time* using multiple measures (NCATE, 2000; Takona, 2003). Teachers are usually require to pass some type of basic skills tests such as Praxis I or Praxis II as one required data point upon program completion (Selk, Mehigan, Fiene, & Victor, 2004). Other states have created their own tests to use instead of the Praxis exams. The passing rates of students on these exams can serve as one of the measures *over time*.

The other measure *over time* that occurs during course completion is more complicated. In previous years, perusal of course syllabi was sufficient to show outcomes of teacher education programs. As time went on, standards were changed and departments were required to collect actual assessment evidence of each departmental goal (Fetter, 2003). Thus, there has been a major shift in the assessment of students in education programs. Departments often opt to have students compile a portfolio of critical tasks or assessments for faculty review in order to document objective acquisition. The portfolio assessments are to reflect knowledge gained on all departmental objectives/learning outcomes. Programs now must show that candidates have mastered the selected outcomes and that these outcomes have a positive impact on learners (Fetter, 2003).

Klecker (2000) suggests that there are three major expectations for pre-service teacher portfolios including: providing more meaningful and valid indictors of what pre-service teachers know and can do, enhancing both teaching and learning and providing useful assessment information. Education faculty must determine the assessments contained in the portfolios. Portfolios must include products that clearly demonstrate that the candidate can perform required outcomes, not just exposure to a concept in a course (Fetter, 2003). Assessment entries in portfolios may include written work such as reports, term papers, graded tests, assignments and lesson/unit plans. Other entries may include artwork, lists of conferences, letters from parents, notes from students and video recordings of teachings (Takona, 2003). Products may come from multiple sources such as course work, field experiences and volunteer work. The products must be connected to the program outcomes as established by the conceptual framework (Fetter, 2003).

Other issues include who determines the content of the portfolio and what should be included. Some programs are quite prescriptive whereas others are more student-oriented. The type is determined by the purpose of the portfolio (Dougan, 1996). Accordingly, if the portfolio is to show mastery of content, then the department faculty should choose the products (Dougan, 2003). Stakeholders in the teacher education programs need to determine a specific set of assessments related to the program even if multiple sections are taught by different instructors (Fetter, 2003). These critical assessments need to have consistent assignment descriptions and rubrics to provide consistency in scoring.

Purpose of the Article

The purpose of this article is to present three easy to implement methods for augmenting the faculty consensus method which may be associated with better assuring the alignment of department goals and assessment tasks.

Need for Valid Critical Portfolio Tasks

The idea of the faculty choosing the assessment content and the inclusion of valid tasks are closely related. It is imperative that the chosen critical tasks and assessments are held to the same standards as most measurement systems which require estimations of reliability and evidence of validity (Ghiselli, Campbell, & Zedeck, 2001; Mehrens, 2001; Klecker, 2000). Reliability estimates allow for an examination of consistency of scoring by professors on critical assessments. Evidence of the content validity allows the department to suggest that the critical assessments represent the state objectives well. This evidence is of primary importance to any assessment system because there are inferences made on mastery based on assessment ratings. That is, students who score high enough on critical assessments are considered to have demonstrated mastery of those objectives. This inference is only true if the task/assessment represents the objective in a meaningful way. Providing evidence of the validity of critical tasks in students' portfolios may be done in several ways and to varying degrees.

The *Standards for Educational and Psychological Testing* (AERA, 1999) is a joint publication of the American Educational Research Association, American Psychological Association and the National Council on Measurement and it suggests that the most common method of providing evidence of content validity for any test or assessment is to have content area experts rate the degree to which each test item represents the objective or domain. These standards may be applied to portfolios. The items are like the critical tasks and the domain is represented by the state objectives. The validity question with portfolios becomes how well each critical task represents the mandated goal. The alignment of assessment task and state goal is of key concern. There are several ways to assure the alignment between of assessment tasks and objectives that vary in degrees of certainty.

A fairly typical way of trying to provide evidence of assessment task validity is to use faculty members as content/domain experts. This method would require faculty to work collaboratively and agree upon the representativeness of each critical task. Individual faculty could write critical task/assessment descriptions associated with each objective. The faculty can then discuss each task and come to consensus on the tasks' ability to represent the objective. The method is informal and fairly simple to implement but consensus can be difficult. Sometimes there are more critical tasks identified than needed per goal so departments need to select the most valid or representative assessment tasks to include in student portfolios. If consensus is difficult or there are more tasks identified than needed or if faculty just want to be extra certain of task validity, faculty may opt to use one of the following easy methods to evaluate task validity/representativeness.

Methods for Providing Additional Evidence of Portfolio Task Validity
Q-Sort Method

There is another step education faculty could take that may create a more organized collaborative discussion and perhaps be associated with more valid results. Again, classical measurement literature has suggested the use of a Q-sort as a method for looking at representativeness (Crocker & Algina, 1986). Q-sorts force a ranking of items by content experts. They have had wide application and could easily be adapted for use here. The method would require more effort on the faculty's part but may be associated with increased validity. The process could be broken down into the following steps:

1. Faculty would have an objective and a list of possible critical assessments that could be indicative of the objective.
2. The number of assessment tasks used should exceed the number of tasks needed to represent the domain or objective.
3. Each faculty is asked to rank order each assessment by representativeness to the objective.
4. Data is collected for each individual assessment task and mean rankings calculated.
5. Assessment tasks with the highest rankings are the tasks selected for use.

This method serves to formalize the dialogue between faculty members by forcing a ranking from each faculty member as to the representativeness of each critical task. The rankings may require a more careful consideration than dialogue. There may be more representativeness certainty with dialogue and rankings than with dialogue alone. The more certainty that assessments tasks represent state goals, the more valid the assessment tasks are.

Lawshe's Content Validity Ratio

There is another method borrowed from the classical measurement literature which has applications for the evaluation of portfolio task representativeness. Lawshe (1975) created a Content Validity Ratio (CVR) that is used to gauge the content validity of items on an empirical measure. In this approach, a panel of Subject Matter Experts (SMEs) is asked to indicate whether or not a measurement item in a set of items is "essential" to the operationalization of a theoretical construct. "Essential" items or assessment tasks are ones that best represent the goal and are desired. Faculty members may be used for SMEs. The measurement item in this case is one of several possible portfolio tasks and the construct is the goal. For example, the portfolio assessment task may require the pre-service teacher to construct a traditional test following item writing guidelines and the state goal is "assessment". The question to the SMEs becomes to what degree, on a scale of one to five with five being very essential and one being not essential at all, is the construction of a traditional test to "assessment" in the classroom. There could be another possible portfolio task in the set of items which requires the candidate to

write varying levels of objectives according to Bloom's Taxonomy. Again, the question would become how essential do the SMEs rate this task of objective writing to assessment. The two assessment tasks would have varying ratings of "essentialness" but the best and most valid assessment task would be the one with the highest CVR because the ratio indicates the proportion of "essential" ratings.

The following is the ratio used after collecting "essential" responses:

$$CVR = (2n_g / N) - 1$$

Where n_g is the number of SMEs who think the item is good and N is the total number of SMEs. Again, the SMEs should be rating the items/portfolio tasks in terms of representativeness and essentialness to the goal. One can infer from the equation that the CVR takes on values between -1.00 to +1.00, where a CVR = 0.00 means that 50% of the SMEs in the panel size of N believe that the portfolio task is essential thereby valid. Lawshe has further established minimum CVRs for varying panel sizes based on a one tailed test at the $a = 0.05$ significance level. For example, if 25 SMEs make up the panel, then measurement items for a specific task whose CVR values are less than 0.37 would be deemed as not essential enough and deleted from use. Faculty could submit several assessment tasks to consider representing a single particular goal and use the ones with the highest CVR as evidence of the content validity of their assessment tasks. This method would provide the department with quantitative data about the validity of each accepted assessment task being used to measure goal mastery.

There is a third procedure that can serve to augment either of the above mentioned methods. This method is more time consuming but could provide field based evidence of validity which may be optimal because of the high stakes nature of portfolio tasks. The field based approach is described below.

Field Testing Procedure

Departments could conduct a field action study to assess how well each critical task previously delineated by university faculty represents the goal. A department may want to consider using people employed in the field as content experts. In the case of education departments, teachers could be considered as content experts. It is thought that they may be good judges of how well assessments represent the objectives in a real life way.

A sample of teachers could be given a brief description of each of the objectives and a list of several possible associated critical tasks for each objective already identified by university faculty. The teachers would be instructed to read each task and rate each the representativeness of each critical assessment to the associated objective on a Likert scale. The survey could include ratings of one to five, with a rating of five indicating the most representative of the objective and a rating of one indicating the least representative of the objective. High ratings would be associated with valid assessment tasks. The data allows for calculation of mean scores for each critical assessment task. The low means suggest that the respective assignments be reevaluated in terms of their relationships to

their intended objectives. Individual faculties could decide acceptable mean benchmark standards and review any means that fall below that level.

Discussion

The intent of this article is to suggest that a triangulation of validity evidence be considered when making decisions about critical assessments for student portfolios. Faculty discussion and informal consensus may not be enough on their own when dealing with high stakes assessment that is being overseen by accrediting agencies. The intent is not to force formal experimental research but rather to consider use of an informal strategy to augment collaborative decisions made by faculty acting as content experts.

The suggested action study used teachers in the field as a second source of content area experts. Another group of content area specialists may be other professors in the state. A department may want to ask professors from other universities to rate the degree of representativeness of each critical task to the state objective. Departments could also collect data as to what type of tasks are varying universities using and compare the critical assessments by doing an informal content analysis. There are different methods available for education departments to use when trying to provide evidence of the validity of their portfolio assessments. The important factor is the recognition of the need to extend beyond typical faculty consensus in situations where faculty consensus may be difficult, more tasks are identified than needed or when departments feel the need to confirm consensus because of the high stakes nature of portfolios.

References

Dougan, A. (1996). Student assessment by portfolio: One institution's Journey. *The History Teacher, 29*(2), 171-178.

Klecker, B. (2000). Content validity of pre-service teachers' portfolios in a standards-based program. *Journal of Instructional Psychology, 27*(1), 35-39.

Ghiselli, E., Campbell, J., & Zedeck, S. (1981). *Measurement theory for the behavioral sciences.* New York: W.H. Freeman and Company.

Mehrens, W. (1992). Using performance assessment for accountability purposes. *Educational Measurement: Issues and Practice, 11*(1), 3-9.

Selk, M., Mehigan, S., Fiene, J., & Victor, D (2004). Validity of standardized teacher test scores for predicting beginning teacher performance. *Action Teacher Education, 25*(4), 20-29.

Takona, J.P. (2003). *Development for teacher candidates.* College Park, MD. ERIC Clearing House on Assessment and Evaluation. (ERIC Document Reproduction Services No. 481816).