Validity and Reliability of Portfolio Assessment of Competency in a Baccalaureate Dental Hygiene Program

Cynthia C. Gadbury-Amyot, Ed.D.; Juhu Kim, Ph.D.; Richard L. Palm, Ed.D.; G. Edward Mills, Ph.D.; Elizabeth Noble, Ph.D.; Pamela R. Overman, Ed.D.

Abstract: This study examined the validity and reliability of portfolio assessment using Messick's unified framework of construct validity. Theoretical and empirical evidence was sought for six aspects of construct validity. Seven faculty raters evaluated twenty student portfolios using a primary trait analysis scoring rubric. A significant relationship (r = .81-.95; $p < .01$) between the seven subscales in the scoring rubric demonstrates measurement of a common construct. There was a significant relationship between portfolios and GPA (r = .70; $p < .01$) and the NBDHE (r = .60; $p < .01$). The relationship between portfolios and the Central Regional Dental Testing Service (CRDTS) examination was both weak and nonsignificant (r = .19; $p > .05$). A fully crossed, two-facet generalizability (G) study design was used to examine reliability. ANOVA demonstrated that the greatest source of variance was the scoring rubric itself, accounting for 78 percent of the total variance. The smallest source of variance was the interaction between portfolio and rubric (1.15 percent). Faculty rater variance accounted for only 1.28 percent of total variance. A phi coefficient of .86, analogous to a reliability coefficient in classical test theory, was obtained in the decision study by increasing the subscales to fourteen and decreasing faculty raters to three. In conclusion, the pattern of findings from this study suggests that portfolios can serve as a valid and reliable measure for assessing student competency.

Dr. Gadbury-Amyot is Director, Division of Dental Hygiene, University of Missouri-Kansas City; Dr. Kim is Assistant Professor, Counseling, Education Psychology, and Exercise Science, University of Missouri-Kansas City; Dr. Palm is Assistant Professor, Division of Urban Leadership and Policy Studies in Education, University of Missouri-Kansas City; Dr. Mills is Dean, College of Education, Pacific University, Forest Grove, Oregon; Dr. Noble is Associate Professor, Division of Urban Leadership and Policy Studies in Education, University of Missouri-Kansas City; and Dr. Overman is Assistant Dean of Academic Affairs, School of Dentistry, University of Missouri-Kansas City. Direct correspondence and requests for reprints to Dr. Cynthia C. Gadbury-Amyot, Associate Professor and Director, Division of Dental Hygiene, School of Dentistry, University of Missouri-Kansas City, 650 E. 25th Street, Kansas City, MO 64108; 816-235-2050 phone; 816-235-2157 fax; amyotc@umkc.edu.

Key words: portfolio assessment, validity, reliability, generalizability theory

Submitted for publication 5/8/03; accepted 7/15/03

Competency has been defined as the skills, understanding, and professional values of an individual ready to begin practicing independently.[1] More specific to dental and allied dental practice, competencies are outcomes of classroom and clinical training combined with experience. In the educational environment, faculty members are traditionally charged with the responsibility for defining and evaluating competence. While the responsibility for defining competencies should lie with the faculty, one of the hallmark characteristics of a competent individual is the ability to accurately assess his or her own competence.[1,2] If we are to graduate competent dental and dental hygiene students, there should be opportunities throughout the curriculum for students to develop self-assessment. One method that has been used in other disciplines is authentic assessment or performance assessment in the form of portfolios.

Portfolios are a focused purposeful collection of student work that documents evidence of traditional and nontraditional sources of student learning, progress, and achievement over time.[3-7] Because they contain longitudinal information, portfolios can be evaluated for degree of improvement as well as for overall quality. Beyond the presentation of examples of students' work, portfolios contain reflective statements written by the student. In these written reflections, students need to demonstrate to the reader the value of their work examples included in the portfolio and what the examples demonstrate about their intellectual growth. Self-reflection requires analysis and synthesis of thought and action, encouraging active involvement and a sense of ownership in the development of the portfolio and of one's own learning. In discussing portfolio pedagogy, Yancy states that it is in this reflective exercise that portfolios become more than a mere scrapbook.[7]

Several barriers to implementation of portfolio assessment have been identified, with time being one of the largest.[8,9] Portfolio assessment demands a large time commitment from both faculty and students. Another barrier relates to power and control in the educational environment.[10] Educators have often been reluctant to share control with students when it comes to evaluation. Likewise, students have been reluctant to assume responsibility for their own assessment because they have not been accustomed to being accountable for their own learning. The process of increasing metacognition, self-awareness, and responsibility for one's own learning is largely atypical in American education. In its simplest form metacognition is the process of thinking about one's thinking, as well as the individual's ability to assess and regulate the course of his or her thinking in order to achieve specified goals. Instead students frequently do only what they need to get by, and instructors coerce them with the threat of poor grades. Clearly, portfolio assessment requires a paradigm shift for educators and students alike.

Another barrier to implementation of portfolio assessment relates to validity and reliability. The overarching question of whether portfolio assessment is psychometrically sound and can produce psychometrically defensible results is a question being asked across the country and at all levels of education.[11,12] Despite strong theoretical and practical implications for portfolio assessment in dentistry and dental hygiene, there are few studies investigating their use. In instances where researchers reported using portfolio assessment, they did not report validity and reliability evidence.[3,8] Until it can be demonstrated that portfolios are a valid and reliable measure of student learning and competency, widespread acceptance by faculty is unlikely. The purpose of this study was to investigate the validity and reliability of portfolio assessment of competency in a dental hygiene program using Messick's unified framework for validity.[13,14]

Methodology

The Standards for Educational and Psychological Testing define validity as "the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (p. 9).[15] This definition considers validity as a unified concept and is based in part on Messick's theory of validity.[13,14,16] Messick argued that validity is a unitary concept and proposed that validation of performance assessment, as with all other forms of assessment, should incorporate six aspects of construct validity: content, substantive, structural, generalizability, external, and consequential. Reliability has been defined as the measure of the degree of consistency in examinee scores over replications of a measurement procedure.[17]

Validity Model

Using Messick's unified framework for validity, six aspects of construct validity were examined both theoretically and empirically.[13,14] An explanation of each of the six aspects follows. First, Messick's content aspect of construct validity analyzes the content of the tasks, the curriculum, and the domain theory. Typically, this part of construct validation relies on expert judgments about the boundaries of the construct, the curriculum, and the skills and content measured by the tasks. Second, the substantive aspect of construct validity looks for evidence that demonstrates that the tasks in the portfolio assessment lead examinees to engage in the intended cognitive processes, minimizing construct-irrelevant factors. Third, the structural aspect of construct validity addresses the adequacy and appropriateness of scoring and scaling. Fourth, the generalizability aspect of construct validity deals with the replicability or reliability of results across multiple levels of facets of the assessment procedure. Fifth, the external aspect of construct validity explores the relationship of test scores (portfolio scores) to variables external to the test. Presumably persons who score high on portfolio assessment of competency should also score high on other presumed indicators of student competency. And finally, the consequential aspect of construct validity examines the degree to which portfolio assessment had both the intended positive effects and plausible unintended negative effects.

Student programmatic portfolios were implemented in the dental hygiene program by the primary investigator as a nontraditional assessment measure of competency. Implementation required an extensive review of the curriculum to determine how each course related to one or more of the program competencies, as well as what assignments, projects, or evaluations served as evidence of those competencies identified for each course. A table was then developed that included each of the program competencies and items that would serve as evidence of

attainment (course assignments, projects, etc.) for inclusion in student portfolios. Faculty identified items they felt were essential for inclusion in all student portfolios and provided additional examples of items that students might want to consider for inclusion. Students used this table as a guide in the construction of their portfolios. Details on implementation of portfolio assessment, the eight program competencies, and an example of the competency table have been previously published.[8]

Subjects

The sample for this study consisted of a subset of student portfolios. This subset consisted of portfolios of ten dental hygiene students with the highest overall GPA (five from the Class of 2001 and five from the Class of 2002) during their tenure in the program, along with ten students with the lowest overall GPA (five from the Class of 2001 and five from the Class of 2002) during their tenure in the program. All twenty participants were full-time, female students. The mean age and GPA for the ten highest ranked students was 24.5 (*SD* = 4.5) years of age and 3.79 (*SD* = .16) GPA. The mean age and GPA for the ten lowest ranked students was 28.2 (*SD* = 7.7) years of age and 2.76 (*SD* .12) GPA. The ethnic breakdown of the group was fifteen Caucasians, three Asians, one African American, and one Hispanic (Table 1).

Instrumentation and Data Analysis

It has been suggested that scoring rubrics work well as an evaluation instrument for performance assessment tasks such as portfolios.[18,19] A primary trait analysis scoring rubric was developed by the principal investigator, based on a review of the relevant literature, for the purpose of scoring student portfolios. The scoring rubric and details on development have been previously reported.[8] The rubric contains a total of seven traits with thirty-five individual components that serve to capture the essence of each trait, with a four-point Likert scale (1 = no evidence of the trait, 4 = complete evidence of the trait).

Each of the thirty-five components was evaluated separately and summed by trait to create seven subscale scores. These subscale scores served as variables for both generalizability and correlational analyses, with portfolios as nontraditional measures of student competency. Traditional assessment measures of student competency included: grade point average (GPA), the National Board Dental Hygiene Education (NBDHE), and the Central Regional Dental Testing Service (CRDTS) examination. In addition, all thirty-five components were summed to compute an overall score that served as a variable for additional correlational analyses.

To explore consequential validity, an open-ended survey was used to elicit student feedback on the perceived value of the assignment and student perceptions of attainment of program competencies after completion of the assignment. This survey has been previously published.[8] Verbatim comments from the survey were used to examine consequential validity in terms of the intended positive effects and plausible unintended negative effects of portfolio assessment as described by the students themselves.

For purposes of this study, all seven full-time faculty evaluated all twenty student portfolios using the scoring rubric. This study was approved by the University of Missouri-Kansas City Social Sciences Institutional Review Board.

Results

The question of validity and reliability of portfolio assessment was explored through the examination of the degree to which evidence and theory supported the interpretations of portfolio assessment within Messick's framework for construct validity. Theoretical and empirical evidence of the content aspect of construct validity was first examined. Content validation relies on expert judgments about the boundaries of the construct, the curriculum, and the skills and content measured by the tasks. The experts for this study consisted of seven full-time fac-

Table 1. Demographic characteristics of participants (N = 20)

Characteristic	*n*	percent	*M*	*SD*
Age of students with highest GPA	10	50	24.5	4.5
Age of students with lowest GPA	10	50	28.2	7.7
GPA for highest-ranked students			3.79	.16
GPA for lowest-ranked students			2.76	.12
Ethnicity				
Caucasian	15	75		
Asian	3	15		
African American	1	5		
Hispanic	1	5		

ulty in the division of dental hygiene who are responsible for the majority of didactic and clinical instruction in the program.

All seven full-time faculty were involved in the development of eight program competencies as the result of a comprehensive review of the curriculum. Each course was then evaluated in light of the program competencies to determine how courses in the curriculum contribute to student development in the eight program competencies. Course content including tests, projects, and assignments were evaluated to ensure that they contributed to students' attainment of the eight program competencies upon completion of the curriculum. Finally, all faculty participated in the development of the seven traits used for evaluation of portfolios. Consequently, there is strong theoretical evidence for Messick's content aspect of construct validity as a result of the extensive involvement of the faculty in the curriculum, portfolio development, and determining boundaries for the domain (dental hygiene student competency). Empirical evidence was sought through examining the intercorrelations between the seven traits or subscales (Table 2). A significant ($p < .01$) relationship was found between each of the subscales in the portfolio scoring rubric used for measuring student competency. Cronbach's alpha (α), or the internal consistency of the subscales in the portfolio scoring rubric, were calculated. Cronbach's alpha ranged from 0.81 to 0.95 on the seven subscales. While Cronbach's alpha is a measure of reliability, it helps to strengthen the argument that a common construct was measured by the scoring rubric.

Next, theoretical evidence of the substantive aspect of construct validity was examined through an analysis of the literature on competency. Messick contends that the substantive aspect should demonstrate that the processes used in completing the tasks are representative and relevant to the processes that constitute the construct of interest.[13,14] Researchers and experts on competency-based education posit that one of the hallmark characteristics of a competent individual is the capacity to accurately assess or self-evaluate competence.[1,20,21] Faculty believe that student involvement in the selection of portfolio content and subsequent self-reflection and self-evaluation of portfolio entries as they relate to attainment of program competencies serves to develop students' ability to assess or self-evaluate their competence. Therefore, it would appear that the process of self-evaluation used in the construction of portfolios is relevant to the construct of student competency. Construct-irrelevant variance refers to the degree to which test scores are affected by processes that are extraneous to its intended construct. An example might be a highly anxious reaction to a test situation. Student portfolios contain many examples of student work produced throughout their two years in the program, where multiple faculty have provided evaluations. This longitudinal approach to assessment is more likely to rule out construct-irrelevant variance than one-time measures of assessment in which an evaluation of competency is being sought. Consequently, a good argument can be made that there is sufficient theoretical evidence to support Messick's substantive aspect of construct validity in relation to portfolio assessment of student competency.

Theoretical evidence of the structural aspect of construct validity was examined through an analysis of the literature on performance assessment. Structural validity examines the scoring system as it relates to the construct domain. Messick states that the scoring method should be consistent with the construct domain.[13,14] Educational assessment experts suggest that primary trait analysis (PTA) scoring rubrics are effective scoring systems for performance assessment measures, in this case portfolios.[19,22] This was the scoring system employed in the evaluation

Table 2. Intercorrelations and coefficient alphas for the seven subscales in the portfolio scoring rubric as content validity evidence

	Growth	Attainment	Self-Evaluation	Lifelong Learning	Organization	Creativity	Communication
Growth	**.81**						
Attainment	.71	**.95**					
Self-Evaluation	.78	.63	**.90**				
Lifelong Learning	.78	.54	.79	**.92**			
Organization	.77	.69	.72	.67	**.92**		
Creativity	.81	.78	.79	.74	.81	**.92**	
Communication	.78	.61	.76	.73	.84	.84	**.91**

Note: Coefficient alphas are presented in boldface along the diagonal. All correlations are significant at $p < .01$ level.

of student portfolios. To test structural validity statistically and seek empirical evidence, item analyses were conducted on the thirty-five items in the PTA to determine how the items conformed to the subscales. There was a significant relationship ($p < .01$) between all thirty-five items in the PTA. Next, correlations between item score and total score for the item's subscale and correlations between item score and total score of the other six subscales were computed (Table 3). In terms of psychometrics, an item should be correlated with its own scale (convergent validity), and it should be correlated more with its own scale than with other scales (discriminant validity).[23] In all but one instance (item 2 in the Growth/Development subscale), items were more highly correlated with their own subscale than with the other subscales. Discriminant validity is based on low correlations of items across subscales. In this study, correlations of items across subscales were overall moderate in strength. Therefore, while a general construct pattern appears to have emerged from this analysis, caution in interpretation must be exercised since item homogeneity is a threat to validity.

Empirical evidence of the generalizability aspect of construct validity was sought through a generalizability (G) study and a subsequent decision (D) study. A G study examines the reliability or replicability of assessment results. The advantage of generalizability theory over classical test theory is that multiple sources of error in a measurement can be estimated separately in a single analysis. The total variance across all facets and interactions between facets can be decomposed into individual components for analysis using ANOVA. For purposes of this study, a fully crossed, two-facet design was used. Shavelson and Webb state that a two-facet design has seven sources of variability: three main effects, three two-way interactions, and one three-way interaction (Figure 1).[24] The three main effect sources of variance for this study were portfolio, rubric, and faculty. The three two-way interactions were portfolio x rubric, portfolio x faculty, and rubric x faculty. Finally, the one three-way interaction was portfolio x rubric x faculty. Estimated variance components and percentage of total variance for each of the sources of variability appear in Table 4.

Results show that the largest variance component, that for rubric (42.8155), accounted for approximately 78 percent of the total variance. The interaction between portfolio and rubric was very small (.6335), accounting for approximately 1 percent of the variance. This illustrates that while the seven subscales within the scoring rubric varied in difficulty level, the relative standing of individual portfolios was maintained across the seven subscales. The variance component for faculty was small (.7005) accounting for 1.28 percent of the total variance, indicating that faculty accounted for little of the variability in portfolio measurement. The variance component for the interaction between rubric and faculty was also small (1.0503, 1.91 percent). These results demonstrate that faculty interpreted the rubrics similarly and were well calibrated. The variance component for the interaction between portfolio and faculty (3.1162, 5.68 percent) was somewhat larger but still only accounted for about 6 percent of the variance in portfolio measurement. So even though faculty used the rubrics in a calibrated fashion, they disagreed to a small degree in the relative standing of portfolios.

Next, estimated variance components from the G study were used in a Decision (D) study to determine the optimal number of facets needed to obtain an acceptable level of reliability when evaluating student portfolios. Table 5 presents the estimated variance components (σ^2), error variance estimations (σ^2_{Rel} and σ^2_{Abs}), and generalizability coefficients (ρ_{2Rel} and φ_{Abs}) for different decision studies, varying the number of faculty and subscales within the scoring rubric. Because portfolios are a criterion-referenced measure, the investigators were interested in Absolute Decisions (φ_{Abs}), where all variance components except the universe-score variance compo-

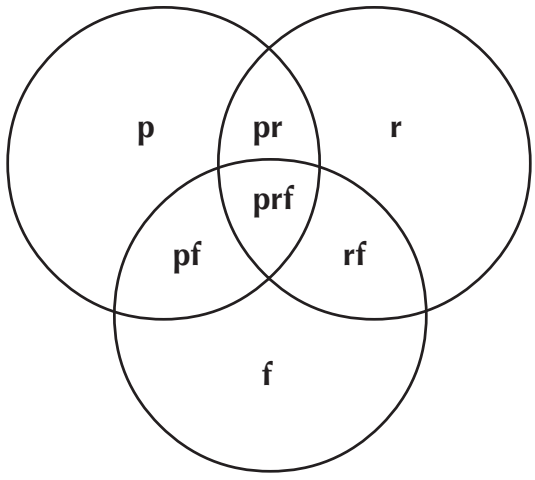

Figure 1. Venn diagram for a two-facet portfolio (p) x rubric (r) x faculty (f) fully crossed design

Table 3. Correlations of each item with its own scale (bold) and with other scales as structural validity evidence (convergent and discriminant validity)

Items	Scales G/D	A	SE	LL	O	C	CM
Growth and development items							
Illustrates continued development and growth over time, i.e., ability to read, analyze, and apply scientific literature in decision making process.	**.84**	.68	.77	.71	.72	.80	.71
Demonstrates increased use of professional language over time.	**.82**	.63	.77	.71	.74	.74	.83
Illustrates heightened professionalism, humanitarianism, and ethical behavior.	**.85**	.69	.61	.62	.68	.69	.63
Provides evidence of professional involvement, i.e., SADHA, Students Take Action, student office of state or local dh associations, etc.	**.75**	.36	.44	.52	.43	.47	.42
Attainment of competency items							
Competent in assessing persons of all ages/stages of life.	.63	**.85**	.52	.44	.59	.67	.51
Competent in dh treatment planning and case presentation.	.68	**.91**	.60	.56	.60	.71	.55
Competent in health education strategies.	.64	**.90**	.58	.47	.60	.67	.53
Competent in provision of preventive and therapeutic dh services.	.56	**.87**	.52	.45	.55	.65	.44
Competent in use of supportive procedures.	.64	**.92**	.56	.48	.61	.67	.53
Competent in infection and hazard control procedures.	.61	**.83**	.57	.49	.62	.72	.58
Competent in management procedures.	.60	**.91**	.57	.43	.61	.69	.55
Competent in community oral health strategies.	.63	**.87**	.53	.46	.67	.69	.57
Self-evaluation items							
Identifies weaknesses.	.64	.49	**.88**	.66	.58	.65	.61
Identifies strengths.	.71	.71	**.84**	.65	.64	.70	.67
Identifies ways to enhance areas of weaknesses and build on strengths.	.67	.48	**.89**	.71	.61	.67	.68
Documents self-evaluation using external evidence.	.72	.55	**.88**	.73	.69	.73	.71
Lifelong learning items							
Contains explanation of student's commitment to lifelong learning.	.72	.54	.74	**.92**	.62	.69	.69
Contains explanation of student's short- and long-term career goals.	.67	.40	.65	**.90**	.52	.59	.57
Student demonstrates the value of lifelong learning to them personally and to the profession as a whole.	.70	.49	.75	**.92**	.61	.70	.67
Able to use information technology to assist in evidence-based decision making (ability to find information for answering questions).	.72	.52	.69	**.85**	.69	.70	.69
Organizational items							
Portfolio design is concise with logical organization.	.71	.68	.61	.59	**.84**	.71	.74
Reflects student's ability to adequately manage information and assemble relevant items to support achievement of competence.	.71	.65	.65	.62	**.93**	.75	.80
Student demonstrates ability to interpret faculty guidelines for development of portfolio with a final product that demonstrates organizational skills.	.68	.55	.66	.64	**.92**	.71	.76
Reflections on items chosen for inclusion are logical interpretations and organized in such a manner as to support the student's claim of competency.	.68	.62	.67	.57	**.90**	.76	.72
Creativity items							
Student demonstrates ability to interpret faculty guidelines for development of portfolio with a final product that demonstrates individual creativity.	.73	.72	.74	.67	.78	**.90**	.73
Reflects personality and characteristics of the student.	.65	.72	.60	.57	.69	**.87**	.63
Demonstrates ability to apply knowledge in creative (out-of-box) situation.	.79	.73	.79	.72	.73	**.93**	.69
Creative application/inclusion of items not required by faculty to meet portfolio guidelines.	.77	.69	.70	.71	.76	**.94**	.69
Communication items							
Introductions and summary present clear and succinct statements.	.71	.59	.71	.68	.76	.72	**.85**
Organizational pattern of entries are logical and easy to follow.	.72	.62	.64	.59	.83	.75	**.85**
Portfolio contents are referred to as documentation to support the points made by the author.	.45	.23	.47	.43	.47	.36	**.72**
There are few errors in grammar or mechanics to distract from the overall presentation of information.	.59	.41	.52	.56	.63	.58	**.78**
Entries are placed in context for the reader.	.73	.53	.70	.66	.73	.68	**.87**
The relationship between the entry and the program competency is clearly linked for the reader.	.67	.60	.67	.61	.74	.65	**.87**
Illustrates ability to transfer knowledge from school into practical application or the work environment.	.70	.59	.71	.71	.72	.68	**.84**

G/D (Growth/Development Scale) A (Attainment Scale) SE (Self-Evaluation Scale) LL (Lifelong Learning Scale)
O (Organization Scale) C (Creativity Scale) CM (Communication Scale)

Table 4. Estimated variance components for portfolio assessment of student competency as generalizability validity evidence

Source of Variation	Sums of Squares	df	Mean Squares	Estimated Variance Components	Percent of Total Variance
Portfolio (p)	4397.45	19	231.445	4.1386	7.54
Rubric (r)	36132.16	6	6022.026	42.8155	78.05
Faculty (f)	859.76	6	143.293	.7005	1.28
pr	779.40	114	6.837	.6335	1.15
pf	2760.65	114	24.216	3.1162	5.68
rf	842.69	36	23.408	1.0503	1.91
prf,e	1642.91	684	2.402	2.4020	4.38

nents (portfolio) contribute to error. Generalizability coefficients, analogous to classical test theory's reliability coefficient, were computed from the error variance components to determine the best conditions for measurement reliability. Analyses demonstrated that portfolio assessment of student competency using seven faculty raters and seven subscales resulted in a phi coefficient of .81. The D study further illustrated that increasing faculty raters beyond three contributed very little to the reliability or dependability of portfolio evaluation. This was to be expected since the G study found that faculty raters (and interactions with faculty raters) contributed little to the variability of portfolio evaluation. Conversely, because the G study identified the rubric as contributing the greatest degree of variability in portfolio scores, increasing the number of subscales within the rubric seemed the next logical step for the D study. As expected, the D study demonstrated that increasing the subscales within the rubric from seven to fourteen and using three faculty raters resulted in a phi coefficient of .86. As a result of these analyses, there appears to be good empirical evidence for Messick's generalizability aspect of construct validity or reliability of portfolio assessment of student competency.

Next the external aspect of construct validity was examined. External validity concerns the relationship of test scores (portfolio scores) to variables external to the test. Messick contends that central to most validation efforts is that persons who score high on the test (portfolios) should score high on other presumed indicators of the construct being measured.[13,14] Correlational analyses were conducted between student portfolios and traditional assessment measures of dental hygiene student competency including the NBDHE, GPA, and the CRDTS examination scores (Table 6). Results show a moderate and significant relationship between portfolios and GPA

Table 5. Decision study for portfolio assessment of student competency (r x f design) as generalizability validity evidence

			G Study	Alternative D Studies							
Source of		n_f =	1	1	2	3	4	7	3	3	3
Variation	σ^2	n_r =	1	7	7	7	7	7	10	14	20
Portfolio (p)	σ^2_p		4.1386	4.1386	4.1386	4.1386	4.1386	4.1386	4.1386	4.1386	4.1386
Rubric (r)	σ^2_r		42.8155	6.1165	6.1165	6.1165	6.1165	6.1165	4.2815	3.0582	2.1407
Faculty (f)	σ^2_f		.7005	.7005	.3502	.2335	.1751	.1000	.2335	.2335	.2335
pr	σ^2_{pr}		.6335	.0905	.0905	.0905	.0905	.0905	.0633	.0452	.0316
pf	σ^2_{pf}		3.1162	3.1162	1.5581	1.0387	.7790	.4451	1.0387	1.0387	1.0387
rf	σ^2_{fr}		1.0503	.1500	.0750	.0500	.0375	.0214	.0350	.0250	.0175
prf,e	$\sigma^2_{pfr,e}$		2.4020	.3431	.1715	.1143	.0857	.0490	.0800	.0571	.0400
σ^2_{Rel}			6.1517	3.1781	.80	.3646	.2107	.0775	.3552	.3508	.3484
σ^2_{Abs}			50.721	4.7738	1.8586	1.3186	1.1296	.9659	.8623	.6476	.5336
ρ_{2Rel}			.40	.57	.84	.92	.95	.98	.92	.92	.92
φ_{Abs}			.0075	.46	.69	.76	.79	.81	.83	.86	.89

Table 6. Correlations among traditional (GPA, NDHBE, and CRDTS) and nontraditional (portfolio) measures of dental hygiene student competency as external validity evidence

	Portfolio	GPA	NDHBE	CRDTS
Portfolio	1.00			
GPA	*.70	1.00		
NDHBE	*.60	*.64	1.00	
CRDTS	.19	.51**	.42	1.00

*Correlation is significant at 0.01 level
**Correlation is significant at 0.05 level

GPA (Grade Point Average)
NDHBE (National Dental Hygiene Board Examination)
CRDTS (Central Regional Dental Testing Service Examination)

(r = .70; p < . 01), and the NBDHE (r = .60; p < .01). The relationship between portfolios and the CRDTS examination was both weak and nonsignificant (r = .19; p >.05). As a result of these analyses, there appears to be good empirical evidence for Messick's external aspect of construct validity of portfolio assessment of student competency.

Finally, the sixth aspect was the consequential aspect of construct validity. The consequential aspect examines the intended and unintended consequences of test (portfolio) use and the impact on score interpretation and use. In real-world applications, the desirable consequences of using a measurement procedure should outweigh the negative consequences of such use.[25] Messick suggests that intended consequences might include changes in the instructional and curricular practices of teachers that lead to better learning environments for students.[13,14] What Messick defines as an intended consequence actually resulted in an unexpected finding during evaluation of student portfolios. A concept that is taught throughout the curriculum pertains to a comprehensive approach to dental hygiene care. During the spring of 2001 when evaluating portfolios, faculty discovered that students almost unanimously identified specific procedures (removal of overhangs, placement of sealants, and application of desensitization medicaments) as adjunctive therapies rather than components of comprehensive care. As a result, faculty discussed changes in instructional and curricular practices aimed at clarifying the concept of comprehensive dental hygiene care.

Further exploration of intended consequences and examination of unintended or negative consequences as defined by Cronbach and Gleser were aided through the analysis of comments from an open-ended survey used to elicit student feedback on the perceived value of developing programmatic portfolios administered to the graduating classes of 2001 and 2002.[25] The overall response rate from the two classes was 73 percent (41/56). Ninety-five percent (39/41) of the students believed that their portfolios demonstrated achievement of the UMKC Division of Dental Hygiene competencies. Verbatim comments included the following: "Faculty comments and feedback are great proof, you can see the growth from paper to paper, treatment plan to treatment plan. . . . even the terminology and thought processes used mature from semester to semester"; "I think it is the self-evaluation that demonstrates achievement but the documents are obviously necessary"; "It helped me to reflect over the past two years as a student and to reflect over my work as a clinician"; and "My portfolio helped me realize that I had achieved competency in these areas."

In response to the survey question pertaining to whether students saw value in developing programmatic portfolios, 76 percent (31/41) responded yes. A sample of responses from this group included the following: "I was reluctant and full of procrastination at first. But when I actually sat down and started to pull everything together to present the big picture, I became excited and proud of the achievements I have made during this program. Value? Yes, I look forward to sharing this portfolio with as many people as I can"; "In the beginning I thought the portfolio was a bad idea but when I started gathering everything I wanted to include I started to value all my experiences. I know I enjoyed my experience but I didn't realize how much I had actually accomplished until I saw it in my portfolio"; "I was really dreading this at the beginning but once I was finished and reading back over my portfolio it really made me proud and emotional"; "It was a great accomplishment knowing that I was successful in not only completing this project, but also completing this program"; and "Allowed me to see that I have learned so much. Shows how much change has taken place in the last two years. Changes in attitude, scholastic skills and thinking skills." Five students (12 percent) reported that they saw no value in the developing portfolios, and four (10 percent) reported there was some value.

One strong theme that came from this group related to the time commitment required for developing programmatic portfolios. Examples of comments included the following: "I spent way too much

time on this"; "With everything else going on this semester, I had to prioritize the assignment. If this were all I had to focus on, I would feel more strongly about the level of value"; "I did not find much value in this experience. I like putting all of my accomplishment in a presentable notebook but I did not find the reflection helpful"; "I thought the portfolios were busy work for us to do before graduation. . . . I was proud of my work afterwards but I still don't think the project has much value"; "The portfolio was a lot of work and I am sure I learned more than I realize. At the time of putting it together it felt more like busy work. Hopefully as I present it to those I will be interviewing with, I will be able to see more value in this project"; and "I know what I've accomplished while I was in school and I don't really feel that I needed to self-reflect on any of it."

Discussion

Validation involves accumulating evidence to provide a sound scientific basis for the proposed score interpretations. A critical consideration in validity testing is the intended use or consequences assigned to test results. For this study, portfolio assessment of student competency is used as a partial grade in a capstone course taught during the students' last semester in the program. Within this context the current investigators believe the evidence and theory extended in this study provide good support for validity and reliability in interpreting portfolio scores. However, in terms of a more high-stakes environment, such as licensing certification, the issue of generalizability must be further studied. The following section is an analysis of the results placed in the context of literature pertaining to the validity and reliability portfolio assessment and to competency-based education.

The literature is clear that faculty expertise, involvement, and commitment in implementing and sustaining portfolio assessment are central to content validity.[16,26] Examination of the results of a study conducted by Kramer and DeMarais on the construct validity of the NBDHE provides some interesting parallels.[27] Results of a principal component analysis of exam items indicated that the exam was measuring one underlying variable or construct that the authors defined as comprehensive dental hygiene. The argument could be made that, much like the Kramer and DeMarais study on the NBDHE, one variable or construct is being captured in the evaluation of portfolios, that of comprehensive dental hygiene or dental hygiene competency.

In relation to the substantive aspect of construct validity, the literature on competency-based education has repeatedly emphasized that competent performance requires the ability to assess and self-evaluate competency and to exercise judgment.[21,22] The following examples provide strong theoretical evidence for the substantive aspect of construct validity in portfolio assessment of competency. At a state level, research into the use of self-assessment in clinical dental hygiene practice demonstrated that a large percentage of the respondents reported consistently engaging in self-assessment.[28] Furthermore, the study participants perceived improvement in quality of care and value to their employer as a result of their self-assessment.

From a national perspective, it is specifically stated in the Accreditation Standards for Dental Hygiene Education Programs that graduates must be competent in the application of self-assessment skills to prepare them for lifelong learning.[29] A strong case can be made that the development of portfolios that contain longitudinal data is much more relevant to the construct of student competency than traditional measures such as clinical licensing examinations where students are tested at one time on one patient. Chambers and Glassman, who have published extensively on competency-based education in dental education, argue that current clinical licensing examinations (traditional measures of student competency) do not test competency, but rather are simulations appropriate for identifying beginners.[22] They state that these examinations are neither authentic nor competency evaluation because of the highly prescriptive nature of the evaluation circumstances and because they fail to measure large parts of the competency domain, especially judgment, understanding, and professional values.

While good theoretical evidence has been presented to support the substantive aspect of construct validity, empirical evidence is needed. One method suggested by Messick is the use of process model studies. This would involve a qualitative method of judging the process used to solve a problem by asking students to solve a problem orally. Students involved in this study are required to do oral patient case presentations for faculty and peers. Case studies are also considered essential entries in their portfolios. A standardized method for evaluating prob-

lem solving during these oral presentations could be developed and analyzed.

In terms of the structural aspect of construct validity, the literature has emphasized that assessment of competency is best scored using a primary trait analysis in the form of a scoring rubric. Both theoretical and empirical evidence of structural validity of portfolio assessment of dental hygiene student competency was demonstrated. However, the small sample size precludes making any strong inferences. The fact that scoring rubrics are suggested in educational literature for performance assessment such as portfolios and that results are similar to the empirical analysis of the NBDHE demonstrates a reasonable trend toward the support of the structural aspect of construct validity. Miller and Linn state that not only should the scoring method be consistent with the construct domain, but also the implementation of rigorous and systematic scoring procedures is crucial to obtaining comparable scores across scorers as further evidence of structural validity.[16] Results from the G study show a high degree of faculty rater reliability.

The generalizability aspect of construct validity was empirically tested using generalizability (G) theory to determine the reliability of portfolio assessment of student competency. Faculty accounted for very little variability or error (1.28 percent) in portfolio measurement. Similarly, the interaction between rubric and faculty was small (1.91 percent), demonstrating that faculty members were consistent and calibrated in their use of the scoring rubric. The fact that very little error was attributed to faculty raters has been reported throughout the portfolio literature in both small- and large-scale studies.[26,30] While rubric accounted for the majority of the variance (78 percent), in classic test theory Cronbach argues that variance in specific factors (between items within persons) is regarded as true variance.[31] So while the subscales varied in difficulty, the integrity of scoring was maintained from portfolio to portfolio as evidenced by the small variance component contributed by the interaction between portfolio and rubric (1.15 percent). In addition, the D study demonstrated that a high degree of reliability could be obtained using three faculty raters and additional subscales. It should also be noted that, because of the small sample size, a larger data set could find that additional subscales are not necessary or fewer than the fourteen indicated in this D study. Certainly if portfolios are to be used in a high-stakes assessment environment, it would be necessary to obtain a higher degree of reliability than the .81 found in this study.

Empirical evidence supports the external aspect of construct validity. Herman and Winters posit that interpretation of portfolio score meaning becomes supported when these scores relate highly to other good measures of the construct.[11] When one considers that a student's GPA is a demonstration of his or her ability over time measured by multiple evaluators in many circumstances, it makes logical sense that portfolios and GPA would be positively related (r = .70).

A moderate relationship (r = .60) was also found between portfolios and the NBDHE. Both validity and reliability data have been reported in the literature in relation to the NBDHE.[27] In contrast, a weak and nonsignificant ($p > .05$) relationship was found between portfolios and the CRDTS examination. Unlike the NBDHE, CRDTS has never published data related to reliability and validity. Much has been written about the validity, reliability, and ethical issues involved in clinical licensure examinations.[22,32-34] Feil et al.[33] point to inconsistencies between student performance at accredited schools and performance on clinical licensure examinations as examples of the lack of validity and reliability of clinical licensure examinations. A national study involving dental hygiene program directors utilized the Delphi process to gain consensus on the best way to determine clinical competence prior to issuing the dental hygiene license.[34] Results showed that the participants believe that dental hygiene clinical competence is best determined through ongoing evaluations in an accredited dental hygiene program.

Results from studies like Feil et al. and Patrick, combined with the results from our study, provide a strong argument for finding alternative methods of assessing student competency. A combination of student performance on multiple measures exhibiting acceptable validity and reliability, such as the NBDHE and portfolio assessment where case studies of the dental hygiene process of care are documented, arguably provides a more valid and reliable assessment of student competency than a one-time clinical licensing examination with questionable validity and reliability. Furthermore, the primary investigator suggests that external examiners could be trained in portfolio assessment, thus providing a method for assessing dental hygiene competency across programs.

Evidence for the consequential aspect of construct validity was drawn from the analysis of student comments on an open-ended survey on the perceived value of programmatic portfolio development. These same positive and negative findings have been reported throughout the portfolio literature.[9,35]

Certain limitations of our study must be acknowledged when considering the implications of its results. The most obvious limitation relates to sample size. For practical reasons it was not feasible to expect all seven faculty to evaluate all student portfolios in the graduating classes of 2001 and 2002 (N = 56). Although the small sample size (N = 20) limits inferences, the pattern and strength of the findings have implications for the study question of validity and reliability of portfolio assessment of competency. Another limitation relates to the issue of academic preparation and the fact that this study was carried out in one dental hygiene program. Currently, there are 266 dental hygiene programs across the country, with the majority of programs located in two-year institutions.[36] Therefore, these results cannot be generalized across all dental hygiene programs.

Future studies should include examination of the validity and reliability of portfolio assessment of student competency in a more representative sample of dental hygiene students. A study to determine whether portfolio assessment is a valid predictor of future dental hygiene practice competency would provide valuable psychometric evidence related to portfolio assessment. Finally, a study examining the perceptions of faculty involved in portfolio assessment of competency of dental hygiene students would provide valuable insight into this alternative evaluation method.

Conclusions

Although the sample limits our inferences to graduates of the UMKC Division of Dental Hygiene, the pattern of findings from this study suggest that portfolios can serve as a valid and reliable measure for assessing student competency. In addition, portfolio assessment holds promise as a possible alternative to current clinical licensure examinations.

Acknowledgments

This project was supported by a fellowship from GlaxoSmithKline and the ADHA Institute for Oral Health. The primary investigator wishes to thank the students from the University of Missouri-Kansas City School of Dentistry, Division of Dental Hygiene Classes of 2001 and 2002 who allowed us to use their portfolios for data collection purposes. We also thank the dental hygiene faculty—Dr. Bonnie Branson, Professor Kim Bray, Professor Lorie Holt, Professor Nancy Keselyak, Professor Tanya Villalpando Mitchell, and Ms. Colleen Schmidt—because this project would not have been possible without their help and support.

REFERENCES

1. Chambers DW. Some issues in problem-based learning. J Dent Educ 1995;59:567-72.
2. Chambers DW. Toward a competency-based curriculum. J Dent Educ 1993;57:790-3.
3. Lusk LT. The portfolio approach/assessing learning in the dental radiology laboratory. Educ Update 1999;18(1):4-6.
4. Isabel JM. Portfolio assessment in a clinical laboratory science curriculum. Clin Lab Sci 1997;10(3):141-4.
5. MacIsaac D, Jackson L. Assessment process and outcomes: portfolio construction. New Dir Adult Cont Educ 1994;62:63-72.
6. Arter JA, Spandel V. Using portfolios of student work in instruction and assessment. Educ Measurement: Issues and Practices 1992;11:36-44.
7. Yancy KB. Teachers' stories: notes toward a portfolio pedagogy. In: Yancey KB, ed. Portfolios in the writing classroom: an introduction. Urbana: National Council of Teachers of English, 1992:12-20.
8. Gadbury-Amyot CC, Holt LP, Overman PR, Schmidt CR. Implementation of portfolio assessment in a competency-based dental hygiene program. J Dent Educ 2000;64(5):375-80.
9. Mullin JA. Portfolios: purposeful collections of student work. New Dir Teaching and Learning 1998;74:79-87.
10. Burch CB. Inside the portfolio experience: the student's perspective. English Educ 1999;32:34-49.
11. Herman JL, Winters L. Portfolio research: a slim collection. Educ Leadersh 1994;52:48-55.
12. Herman J. Large-scale assessment in support of school reform: lessons in the search for alternative measures. Los Angeles: California University, Center for the Study of Evaluation, Center for Research on Evaluation, Standards, and Student Testing, 1997.
13. Messick S. Validity in performance assessments. In: Linn RL, ed. Educational measurement. New York: American Council on Education and Macmillan, 1989:13-104.
14. Messick S. Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into scoring mean. Am Psychol 1995;50:741-9.
15. American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. Standards for educational and psychological testing. Washington, DC: American Educational Research Association, 1999.

16. Miller MD, Linn RL. Validation of performance-based assessment. Appl Psychol Meas 2000;24:367-78.
17. Brennan R. An essay on the history and future of reliability from the prespectives of replications. J Educ Measurement 2001;38:295-317.
18. Palomba C, Banta T. Assessment essentials: planning, implementing, and improving assessment in higher education. San Francisco: Jossey-Bass, 1999.
19. Nitko AJ. Educational assessment of students. Englewood Cliffs, NJ: Prentice Hall, 1996.
20. Evers F, Rush J, Berdrow I. The bases of competence: skills for lifelong learning and employability. San Francisco: Jossey-Bass, 1998.
21. Wiggins G. Assessment: authenticity, context, and validity. Phi Delta Kappan 1993;75:200-14.
22. Chambers DW, Glassman P. A primer on competency-based evaluation. J Dent Educ 1997;61:651-66.
23. Green S, Salkind N, Akey T. Using SPSS for Windows: analyzing and understanding data. Upper Saddle River, NJ: Prentice Hall, 2000.
24. Shavelson RJ, Webb NM. Generalizability theory. Thousand Oaks, CA: Sage Publications, 1991.
25. Cronbach L, Gleser G. Psychological tests and personnel decisions. Urbana: University of Illinois Press, 1965.
26. Naizer GL. Validity and reliability issues of performance-portfolio assessment. Action in Teacher Educ 1997;18:1-9.
27. Kramer GA, DeMarais DR. Construct validity of the restructured National Board Dental Hygiene Examination. J Dent Educ 1997;61:709-16.
28. Fried JL, DeVore L, Dailey J. A study of Maryland dental hygienists' perceptions regarding self-assessment. J Dent Hyg 2001;75:121-9.
29. American Dental Association Commission on Dental Accreditation. Accreditation standards for dental hygiene education programs. Chicago: American Dental Association, 2000.
30. LeMahieu PG, Gitomer H, Eresh JT. Portfolios in large-scale assessment: difficult but not impossible. Educ Measurement: Issues & Practices 1995;14:11-6,25-8.
31. Cronbach L. Coefficient alpha and the internal structure of tests. Psychometricka 1951;16:297-334.
32. VanDam S, Welie J. Requirement-driven dental education and the patient's right to informed consent. J Am Coll Dent 2001;68:40-7.
33. Feil P, Meeske J, Fortman J. Knowledge of ethical lapses and other experiences on clinical licensure examinations. J Dent Educ 1999;63:453-8.
34. Patrick T. Assessing dental hygiene clinical competence for initial licensure: a Delphi study of dental hygiene program directors. J Dent Hyg 2001;75:207-13.
35. Dutt-Doner K, Gilman DA. Students react to portfolio assessment. Contemp Educ 1998;69(3):159-63.
36. Commission on Dental Accreditation. Annual report. Chicago: American Dental Association, 2003.