ED 453 221 TM 032 572

AUTHOR Mushi, Selina L. P.

TITLE Evaluating Validity and Reliability of Classroom Assessments Using Secondary Data.

PUB DATE 2001-04-13

NOTE 30p.; Paper presented at the Annual Meeting of the American Educational Research Association (Seattle, WA, April 10-14, 2001).

PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.

DESCRIPTORS Educational Assessment; Evaluation Methods; Feedback; *Graduate Students; Higher Education; *Reliability; Researchers; *Student Evaluation; *Teacher Made Tests; *Undergraduate Students; *Validity

ABSTRACT

Analysis of secondary data was used as a way to inform the researcher about the trends in her assessment practices over a 4-year period. This was an important initial step in an effort to develop and integrate high-quality classroom assessment tasks and make sense of assessment information for decision making. Scores from 26 groups of graduate and undergraduate education students from 3 universities in the United States were analyzed. Course goals, objectives, and syllabuses were analyzed. Students' backgrounds and group combinations (age, gender, socioeconomic status) were taken into consideration in determining the consistency of specific assessment tasks in providing feedback to the instructor as researcher. The study results provided evidence of high content validity as well as high construct validity of time and nontimed assessment tasks. Concurrent validity among similar assessments tasks was evident. However, the predictive validity of assessment tasks (individual task to the final score) varied depending on whether the assessment task was nontimed (r=0.20 to r=0.41) or timed. Timed assessment tasks were high predictors of the student's performance in the course (r=0.57 to r=0.01). Timed assessment tasks were more reliable (consistent) than nontimed tasks in providing assessment feedback across similar groups and contexts (contextual reliability). Scores from the nontimed assessment tasks fluctuated more from group to group (r=0.04 to r=0.68) than scores from timed tasks. Nontimed tasks probably tapped student skills and strategies that were not retrievable through timed examinations. The study highlights the importance of understanding, documenting, and evaluating assessment practices to better inform decision making at both classroom and program levels. (Contains 2 figures, 3 tables, and 17 references.) (SLD)

EVALUATING VALIDITY AND RELIABILITY OF CLASSROOM ASSESSMENTS USING SECONDARY DATA

Paper presented at the
American Educational Research Association Annual Meeting,
Seattle Washington, April 13th 2001.

Selina L. P. Mushi, Ph.D., Assistant Professor,
Department of Teacher Education, Early Childhood Program,
Northeastern Illinois University

Introduction

In the past decade the debate on classroom assessment has focused more on performance assessments and less on paper and pencil tests. This trend arose from concerns about the inadequacy of a single test or a battery of tests to provide a comprehensive picture of students' abilities in a subject area, and more importantly, lack of transfer of the theoretical knowledge to solving actual real life problems despite high scores on tests. By giving students opportunities to actually "perform" a task, the teacher can assess the students' skills in more realistic ways. Measurement and evaluation books now incorporate information on performance assessments (see for example, Kubiszyn & Borich, 1996; Thorndike, 1997; Gallagher, 1993; Oosterhof, 1994; Ward & Murray-Ward, 1999).

As classroom assessment is becoming more and more comprehensive, determining evidence of validity (accuracy) and reliability (consistency) are also becoming more complex. Validity and reliability coefficients are pretty straightforward to compute using paper and pencil test scores. Determining validity and reliability of varied group projects is not easy. Since classroom assessment is supposed to inform decision-making, validity and reliability of the various assessments must be ensured. This study is an analysis of various class assessment tasks and the resulting numeric scores obtained by students on formative and summative course assessments in 26 classes (groups) taught by the researcher in three different universities over a four-year period in the United States.

Conceptual Framework

In this study, classroom assessment is defined as "*selective collection of representative information on students' learning, accurate processing and accurate interpretation of the information for informed decision making*" This definition was developed from reviewing several definitions of assessment from different textbooks and articles and realizing that none of those definitions covered exactly what the researcher wanted the term "classroom assessment" to mean in this study. This definition implies that the information collected and analyzed has to provide a clear indication as to whether the students have achieved effective learning of the targeted skills and concepts, and whether the students themselves realize the learning from their own perspectives and ways of interpretation. In other words, the assessment data have to show if the students have *experienced the shift from a state of not knowing or not being able to perform a skill*, to a state of *knowing* or *being able to perform that skil*l.

Why do we assess? Teachers gather information about students' learning so that they can get a comprehensive, representative picture of the students' learning processes and the skills already acquired. This information is then used to make decisions such as grade retention or promotion, remedial work, further training, job placements etc. The type of decision to be made will determine the type of information to be collected and how it is analyzed (Airasian, 1996; Gallagher, 1998; Stiggins, 1997; Kubiszyin and Borich, 2000; Eby and Martin, 2001; Weber, 1999).

Comprehensive Assessment: Classroom assessments are becoming more and more comprehensive. Some psychological testing books are also incorporating thought questions (see for example Kaplan and Saccuzzo, 1997; Heitzman, 1997). In the four years during which the data were collected, the researcher needed to make decisions about students' mastery of content, application of skills in real classroom situations with children, communication skills, ability to reflect on their own practices and learn from them, creativity, originality, etc. Since paper and pencil tests could do little to achieve all these goals, the researcher used varieties of assessments. The assessments included, but were not limited to, take-home assignments, group and individual projects, class presentations, observing real classrooms and writing case study reports, reading and summarizing research reports, micro teaching, course portfolios, as well as quizzes, tests and examinations. The reason for such diversification of assessments was to collect information from as many perspectives as possible, to tap the different learning styles and preferences of students.

Assessments are authentic if they are relevant, meaningful, and interwoven into the curriculum (Puckett and Black, 2000; Puckert and Black, 1994; Eby and Martin, 2001; Mindes, Ireton & Madrell-czudnowsky, 1996; Bodrova & Leong, 1996; Danielson, 1996). The information collected about every student needs to be very comprehensive, hence as representative as possible of the capabilities of the student being assessed.

To make the final decision about each student's learning in each course, the researcher added together points from the different assignments, determined the percentage of the points obtained to the total points possible, and assigned the letter grade according to the respective university policies. The grades would then inform other decisions at individual, or program level. Figure 1 demonstrates that component scores determine the overall score, which in turn determines the course grade.

Figure 1 about here

Decision-making: The decision to be made has to be a logical step linked to the information derived from the assessments. Logical organization, appropriate analysis and accurate interpretation and reporting of each student's assessment data therefore cannot be overemphasized. There are different methods of pooling together assessment information for decision making (see for example Kubiszyn and Borich, 1994; 2000). Teachers need to be aware of these different methods, select appropriate ones, and assess how well the methods work for their classes.

Research Method:

The researcher analyzed the classroom assessment task descriptions and requirements, as well as the resulting sets of scores (from her files) to study her own pattern of grading to explore relationships among the sets of scores from different types of assessments. The assessment tasks, the test items, the test scores and the resulting course grades were considered secondary data since their primary purpose of assessing students' learning had been completed.

Studying Content Validity

Each task was analyzed in terms of its description, requirements, scoring criteria, and then compared to the respective syllabus and teaching feedback from students, to examine content validity. Test items were reviewed each time before the test was given to the respective students. Item analysis was carried out after each test in order to identify weak test items, which were then eliminated from the test. Aligning the test items and other assessment tasks with: the specific topics on the syllabus, student feedback on their learning of those topics, the scoring criteria, the obtained scores and the standard error of measurement for each set of scores helped the researcher estimate the content validity of the assessments.

Studying Construct Validity

The tasks were also analyzed in relation to relevant theories to examine their construct validity. Since the courses taught in the four years of the study fell in five main categories, i.e., assessment of learning, education of young children, language acquisition, effective teaching in culturally diverse classrooms, and research methods, fundamental theories were reviewed in these five areas. The researcher aligned the objectives of the course to the goals stated in the prescribed teaching standards, and examined the assessment tasks and test items to crosscheck their links to those objectives. By aligning the assessment tasks to objectives, goals, standards, the philosophy of the course and the implied theories, it was possible to determine construct validity of the assessment tasks.

Studying Concurrent Validity

Correlation coefficients of the various quizzes, tests and examinations were estimated using the Pearson Product Moment Correlation Coefficient method. The Corel Quattro-Pro spreadsheet was used to achieve the correlation coefficient estimations. Figure 2 shows one group's set of scores from different class assignments, the maximum possible scores, obtained scores, and a correlation matrix. The correlation coefficients were used to estimate the concurrent validity of the different tasks. By squaring the correlation coefficients the researcher could more closely estimate the amount of variability in one task that could be explained by the variability in another task.

Figure 2 about here

Studying Predictive Validity

The focus of the study was not on how well single assignments predicted the overall grade in the course. However, the researcher thought it would be interesting to find out if some assessment tasks were better predictors than others, and what the underlying reasons might be.

Studying Reliability of the Assessment Tasks

Sets of scores from similar tasks and tests were correlated using the Quattro Pro Spreadsheet, to estimate the reliability of those tasks across groups of students. Students' scores were matched on the basis of their average performance halfway through the course, including the midterm exam. This approach enabled the researcher to treat each assessment task as an independent measure across groups. The extent to which one assessment task was related to the overall group performance in the course was compared to the extent to which that task was related to overall performance of similar groups. The higher the correlation, the more consistent the task would be.

The researcher did not consider the reliability of the task or test in general terms (which would then, supposedly, apply to any group of students). Rather, the researcher focused on the characteristics of the particular group, and assumed the reliability of the tasks would hold only for groups that could be closely matched. Since there is not a single value that will provide the perfect reliability of an assessment instrument on its own (Thorndike, 1997), the group combinations of students, the students' backgrounds as well as teaching and assessment procedures had to be considered. Lacking a better term to describe this type of consistency for the purpose of the study, the researcher *used contextual reliability* to refer to the degree of consistency of an assessment task across similar groups of students.

Findings

Evidence of Content Validity

Thorough analysis and aligning of the assessment tasks with objectives, goals, and prescribed teaching standards and scoring criteria, showed high content validity. The Standard Error of Measurement values were very low, implying a close match between the obtained scores and the unknown true scores. The researcher did not see the need to adjust the scores.

Table 1 provides a summary of the content validity checks for the different assessment tasks she used in with the 26 groups.

Table 1 about here

As Table 1 shows, the different assessment tasks had content validity of differing extents. For the groups taught in the beginning of the research period, the quizzes had little content validity to the course. Since these beginning of course quizzes were intended to give the teacher/researcher overall preparedness of the students (sizing-up assessment, Airasian, 1996), to take the course, the content was of a general nature. Case studies, reaction papers, research summaries, research papers, class presentations, course portfolios, research proposals (in the few times they were utilized), and class poster sessions indicated moderate-to-high content validity across the 26 groups. Midterm examinations and final examinations were very closely matched to the content of the course, indicating high content validity.

On the whole, the validity checks of all the 26 sets of data indicated that the assessment tasks were matched (M), or closely matched (CM), or even perfectly matched (PM) to the content of the course (see Table 1).

Secondly, an interesting trend emerged from the content validity checks of the assessment tasks. The more recently the group was taught, the closer the match between assessment tasks and the objectives/goals of the course. This trend indicates learning on the part of the researcher, i.e., increased competence in developing content valid assessment tasks.

Evidence of Construct Validity

Examination of the assessment tasks in relation to the guiding theories in each of the five academic areas revealed a strong link, indicating construct validity of the tasks. This type of evidence of validity was easier to determine since the course syllabuses were developed on the basis of the fundamental theories in the respective academic areas.

Evidence of Concurrent Validity

Matching scores from the different assessment tasks indicated varied degrees of correlations among them. The assessment tasks that had mostly moderate to high concurrent validity were: timed examinations (midterm and final exams), take home group tasks (case studies and class poster preparation), and class group tasks (class poster sessions, presentations). Concurrent validity coefficients were consistently low when scores were compared between the following assessment tasks: timed versus non-timed, take-home versus class tasks, group versus individual tasks, and essays versus objective examinations. Table 2 presents this information.

Table 2 about here

Evidence of Predictive Validity

Predictive validity coefficients between the first quizzes and the final score were consistently low (from .2 to .4). This means the students "grew" considerably in unpredictable ways during the course. The midterm exams and final exams were found to be strong predictors of the total score in the course (correlation coefficients between 0.57 and 0.91). One explanation of this high predictability could probably be that students prepared for the exams in similar ways. Examination anxiety could be another factor that kept students at their relative rankings across timed examinations.

Evidence of Reliability of the Assessment Tasks

Consistency of the different assignments over time was of great interest to the researcher. Halfway through the course students had relatively stabilized in achievement. Midcourse ranking of students helped the researcher in matching groups' scores in terms of timed and non-timed, individual and group assignments. For each timed examination, the researcher developed alternative exams and through item analysis the examinations were consistently "cleaned" to get rid of any flaws that might interfere with the quality of the testing process.

Most of the reliability coefficients of the timed exams were considerably high (between 0.73 and 0.86), although there were a few moderate and one low coefficient of 0.2. This means that students kept their relative positions in their groups. However, some of the groups of students were more similar than others.

The non-timed and group assessment tasks were less consistent in determining student performance across groups. Take home group tasks had low or no correlation. This could be explained by the possibility that students learned to work together in the similar ways and scored about the same. The relatively higher correlation coefficients of the take-home individual assignments strengthen the possibility that working together in class and outside class became a factor in obtaining similar scores across these assignments, hence low correlations in group assignments. As such, listing the assessment tasks in order of consistency across groups, timed exams were the most consistent assessment tasks, followed by individual class assessment tasks, individual take-home assessment tasks, group class assessment tasks, and finally group take-home assessment tasks. Table 3 summarizes this information.

Table 3 about here

Discussion

Assessment of learning is useful only when it yields, to a good extent, a representative picture of what the individual student has learned. To achieve a representative picture of student learning, assessment tasks must be closely related to the content based on sound theories, and must be consistent in determining if effective learning of the targeted skills has occurred. Valid and reliable assessments are more likely to provide a representative picture of each student's learning, than are invalid and unreliable assessments.

In this study, timed assessment tasks (tests and exams) seemed to be more consistent in measuring students' learning. They had slightly higher content validity and they were also better predictors of the final score in the course. However, since teachers and evaluators cannot depend on timed exams and quizzes only, there is need to maximize the effectiveness of non-timed, take-home assignments, and use them alongside the timed ones, so that the final grade awarded is closely representative of the individual students' acquired learning in the course. Some ways of making the non-timed and take-home assignments more effective feedback mechanisms, as employed by the current researcher, include: ensuring content validity of the assignments, providing formats for doing the assignments, providing and discussing scoring criteria before hand, and having the students talk about their own work - before scoring is done. Giving students a chance to talk about their projects or take-home papers may highlight important points that might bypass the instructor's/scorer's attention, or that might be misinterpreted due to cultural and language differences. Although take-home and group assignments may not yield high correlations, they will be reliable in the sense that each time they are used they will achieve maximum assessment information about that particular student, and not necessarily the student's ranking within the group.

It is possible that in some cases a student may get help from family members or friends and do well on a take-home assignment one time, and not get the help another time, and do poorly. This could be the cause for the inconsistency of scores as observed in this study, with regard to take-home and group assignments. Fluctuation of scores on take-home assignments could also be due to difficulty levels of those tasks and how much the specific tasks appealed to individual students. Similar group performance (therefore low correlation) indicated possibilities that students helped each other; or low achieving students agreed with high achievers, or high achievers did not do perform at their best. This possibility of helping each other, or under-performing, underscores the necessity to give timed, non-timed, individual and group assignments to students.

While cooperative learning and group projects are a wonderful way of learning, the need to estimate individual capabilities as closely and as validly as possible cannot be overemphasized. With regard to student teachers in particular, once one graduates from a teacher preparation program, the main assumption is that the graduate is ready to take classroom teaching responsibilities and be accountable for students' learning. Cooperation will be a highly desired addition to one's individual competence.

Summary of Emerging Trends and Suggestions for Classroom Teachers

When teaching and assessment are closely interwoven together, the validity of the assessment tasks is enhanced. Both the teaching and the assessment tasks match the same targeted skills and knowledge. When assessment is treated as a final stage of a teaching period, the match between what is taught and what is assessed may be jeopardized. The following are important trends directly and/or indirectly arising from this self-study. The researcher learned important lessons from both the research content and from carrying out the research process itself. These might be useful to other classroom teachers interested in learning more about their assessment practices:

1. Both timed and non-timed assessment tasks were closely related to the target content, hence content validity. Construct validity was confirmed through examination of relevant theories. Concurrent and predictive validity of the different assessment tasks differed considerably, indicating that different assessment tasks measure group performances differently.

However, put together, the assessment tasks measured students' learning more accurately than would timed or non-timed assessment tasks alone.

2. Timed assignments such as quizzes, tests, and examinations, were more consistent in measuring the overall learning in the course, than were the non-timed assignments. This does not mean that timed assessments are inherently better than non-timed assessments. Tests and examinations must be carefully developed and utilized according to testing principles. Awareness and use of professional guidelines in developing quizzes, tests and examinations are the only means to ensure high quality test items that target the intended knowledge and skills without trivializing them. Competence in item writing and item analysis is a pre-requisite for developing and using these types of measures. Haphazard item writing will result in flawed tests with low validity and reliability. Quizzes, tests, and exams are powerful tools of assessment when utilized with trained expertise.

3. A combination of both timed and non-timed, individual and group, take-home and in-class assessment tasks helped the researcher to separate specific areas of the targeted content that could best be assessed using a certain mode of assessment. Pooling together the different assessments consistently indicated that students were given opportunities to demonstrate learning using different means (tasks). Each student's overall

grade for the course was therefore highly representative of the student's achieved learning. Obtaining assessment information from as many perspectives as possible helped accommodate students' different learning styles and modes of expression and talents that might not have been captured by a single assessment tool. The high correlations among timed assessment tasks imply that students kept their class ranks in certain areas (possibly little room for creativity), and fluctuated in other areas (more room for creativity).

4. The researcher realized that providing students with as much information as possible at the beginning of the course helped improve validity of assessments. During the course students became increasingly aware of what was expected of them in doing the assessment tasks. Providing scoring criteria together with the assessments in the beginning of the semester made it easy for the students to understand the tasks and plan their responses.

5. Finally, the researcher strongly suggests consistency, systematicity and self-evaluation in assessment practices. She was able to study her own assessment practices because she kept track of the assessment practices she used with different groups in her four year teaching period. Through modifying assessment tasks and doing item analysis of timed examinations she was able to improve the validity of her assessments. By

studying the consistency of each assessment tool and similar ones in relation to each individual group of students (contextual reliability) she could predict students' performance on a component of the course and take necessary measures before assessment problems occurred.

The researcher suggests that classroom teachers study their own assessment practices to better conceptualize comprehensive assessment and continuously improve validity and reliability. This will lead to more accurately informed decisions about teaching, learning, students' readiness for the job market, and program revisions.

References:

Airasian, P. W. (1996). Assessment in the Classroom. New York: McGraw Hill

Bodrova, E. & Leong, D. (1996). Tools of the Mind: The Vygotskian Approach to Early Childhood Education. New Jersey: Merrill

Danielson, C. (1996). Enhancing Professional Practice. Alexandria: ASCD

Eby, J. W. & Martin, D. B. (2001). Reflective Planning, Teaching, and Evaluation of the Elementary School. A Relational Approach. Third Edition. New Jersey: Merrill

Gallagher, J. D. (1998). Classroom Assessment for Teachers. New Jersey: Merrill.

Heitzman, C. A. (1997). Workbook for Kaplan and Saccuzzo's Psychological Testing.
Principles, Applications and Issues. Fourth Edition. Albany: Brooks/Cole Publishing Company.

Kaplan, R. M. & Saccuzzo, D. P. (1997). Psychological Testing. Principles, Applications and Issues. Fourth Edition. Boston: Brooks/Cole Publishing Company.

Kubiszyn T. & Borch, G. (1996). Educational Testing and Measurement. Classroom Application and Practice. Fifth Edition. New York: harperCollins.

Kubiszyn, T. & Borch, G. (2000). Educational Testing and Measurement. Classroom Application and Practice. Fifth Edition. New York: harperCollins

Mindes, G. Ireton, H. Mardell-Czudnowski, C. (1996). Assessing Young Children. Toronto: Delmar

Oosterhof, A. (1994). Classroom Application of Educational Testing and Measurement. Second Edition. New York: Merrill.

Puckett, B. & Black, J. (2000). Authentic Assessment of the Young Child. Celebrating
Development and Learning. New Jesrsey: Prentice Hall.

Puckett, B. & Black, J. (1994). Authentic Assessment of the Young Child. Celebrating
Development and Learning. New Jesrsey: Prentice Hall

Thorndike, R. M. (1997). Measurement and Evaluation in Psychology and Education.

Sixth Edition. New Jersey: Merrill.

Stiggins, R. J. (1997). Student Centered Classroom Assessment. Second Edition. New Jersey: Merrill.

Ward, A. W. & Murray-Ward, A. (1999). Assessment in the Classroom. Belmont: Wadsworth Publishing Company.

Weber, E. (1999). Student Assessment that Works. A Practical Guide. Toronto: Allyn & bacon.

Abstract

The aim of the evaluation was to combine theory and practice to develop better understanding of classroom assessment in general and assessment practices in particular. Analysis of secondary data was used as a way to inform the researcher about the trends in her assessment practices over a four-year period. This was an important initial step in the effort to develop and integrate high-quality classroom assessment tasks, and making sense of assessment information for decision-making.

The author analyzed scores from 26 groups of graduate and undergraduate Education students in three universities in the United States. Course goals, objectives and syllabuses were analyzed. Students' backgrounds and group combinations (age, gender, socio-economic) were taken into consideration in determining the consistency of specific assessment tasks in providing feedback to the instructor as researcher.

The study results provided evidence of high content validity as well as high construct validity of timed and non-timed assessment tasks. Concurrent validity among similar assessment tasks was evident. However, the predictive validity of assessment tasks (individual task to the final score) varied depending on whether the assessment task was non-timed (r = .20 → r = .41) or timed. Timed assessment tasks (including tests and examinations) were high predictors of the student's performance in the course (r = .57→ r = 0.91).

Timed assessment tasks were more reliable (consistent) than non-timed tasks in providing assessment feedback across similar groups and contexts ("contextual reliability"). Scores from non-timed assessment tasks fluctuated more from group to group (r = .04 → r = .68) than scores from timed tasks (r = .63 → r =.74). Non-timed tasks probably tapped student skills and strategies that were not retrievable through timed examinations.

The study highlights the importance of understanding, documenting and evaluating assessment practices to better inform decision making at both classroom and program levels.

The Course Grade:

- determined by an overall score.

The Overall Score:

- determined by component scores

Component Scores:

- obtained from different types of assignments -
- timed, and non-timed.

Figure 1: Building the Course Grade

9	27	45	88	20	50	88	90	***417***
9	29	47	78	16	53	87	100	***419***
9	22	56	90	19	46	90	105	***437***
8	24	45	95	18	58	95	112	***455***
8	25	34	94	18	53	91	100	***423***
8	21	55	86	20	58	89	100	***437***
7	23	43	85	17	56	88	96	***415***
7	22	54	89	15	55	70	87	***399***
7	27	32	78	17	51	90	98	***400***
7	28	34	80	16	45	90	95	***395***
6	24	26	88	17	54	84	109	***408***
6	20	45	87	17	57	85	100	***417***
6	23	55	96	19	44	90	102	***435***
6	21	46	80	19	58	90	88	***408***
6	19	57	86	20	44	85	90	***407***
5	18	39	85	20	57	78	87	***389***
5	20	46	68	20	50	66	105	***380***
5	15	43	56	17	47	62	90	***335***
5	15	40	55	16	43	59	108	***341***
Max=10	Max=30	Max=60	Max=100	Max=20	Max=60	Max=100	Max=120	Max=500

	Column 1	Column 2	Column 3	Column 4	Column 5	Column 6	Column 7	Column 8	Column 9
Column 1	1								
Column 2	0.718944	1							
Column 3	0.191587	-0.2426425	1						
Column 4	0.519857	0.50928948	0.191768	1					
Column 5	-0.00507	-0.1753617	0.354672	0.254326	1				
Column 6	0.164641	0.13421648	-0.11533	0.360735	0.034884	1			
Column 7	0.642214	0.70062287	0.011241	0.786302	0.232816	0.2964816	1		
Column 8	0.127474	0.10374716	-0.17797	0.039431	-0.09939	-0.081447	0.0971604	1	
Column 9	0.68584	0.57960333	0.297932	0.897259	0.294919	0.3930434	0.872023	0.2559907	

Figure 2: Sample Correllation Matrix

TABLE 1: CONTENT VALIDITY CHECKS

N=No Match = 0 M=Match=1 CM=Close Match=2 PM=Perfect Match=3

Gro up	Quizzes	Case Studies	Reaction Paper	Research Summa-ries	Midterm Exam	Research Paper	Class Present -ations	Course Port-folio	Research Proposal	Class Poster Session	Final Exam	Average Match	Overall Content Validity
1	0	1	NA	NA	3	1	1	NA	NA	NA	3	1.4	M
2	0	1	1	2	3	1	1	NA	NA	NA	3	1.6	M
3	0	2	1	NA	3	1	1	2	NA	NA	3	1.5	M
4	1	2	2	NA	3	1	1	2	NA	NA	3	1.7	M
5	1	2	2	3	3	1	1	2	NA	NA	3	1.9	M
6	0	2	2	3	3	1	1	2	NA	NA	3	1.8	M
7	NA	2	2	2	3	1	1	NA	NA	NA	3	2.0	CM
8	2	2	2	2	3	2	1	2	NA	NA	3	2.1	CM
9	0	2	2	2	3	2	1	2	2	2	3	1.9	M
10	2	2	3	3	3	2	1	2	2	2	3	2.2	CM
11	0	2	3	2	3	2	1	NA	2	2	3	2.0	CM
12	NA	2	2	2	3	2	1	2	2	2	3	2.1	CM
13	0	2	3	2	3	2	1	NA	NA	2	3	2	CM
14	NA	2	3	3	3	2	2	NA	NA	2	3	2.5	CM
15	1	2	3	3	3	2	2	NA	NA	3	3	2.4	CM
16	NA	2	3	3	3	3	2	2	NA	3	3	2.6	CM
17	NA	2	2	3	3	2	2	2	NA	3	NA	2.3	CM
18	2	2	2	3	3	3	2	2	NA	3	NA	2.4	CM
19	2	2	1	3	3	3	2	NA	NA	3	3	2.4	CM
20	2	2	3	3	3	3	2	NA	NA	NA	NA	2.5	CM
21	NA	3	3	3	3	3	3	NA	NA	3	3	3	PM
22	3	3	3	3	3	3	3	NA	NA	3	3	3	PM
23	NA	3	3	3	3	3	2	2	NA	3	NA	2.7	CM
24	NA	3	NA	3	3	3	3	3	NA	3	NA	3	PM
25	NA	3	NA	3	3	3	3	3	NA	3	3	3	PM
26	NA	3	3	3	3	3	3	3	NA	3	NA	3	PM

TABLE 2: CONCURRENT VALIDITY COEFFICIENTS OF ASSESSMENT TASKS

1.0=Perfect Concurrent Validity 0.8=High Concurrent Validity 0.5=Moderate Concurrent Validity Below 0.5=Low Concurrent Validity

GROUP	TIMED EXAMI-NATIONS	TAKE HOME GROUP TASKS	TAKE HOME INDIVID TASKS	CLASS GROUP TASKS	CLASS INDIVID TASKS	TIMED VERSUS NON-TIMED	TAKE HOME VERSUS CLASS TASKS	GROUP VERSUS INDIVID. TASKS	ESSAY VERSUS OBJECTIVE EXAMS
1	0.73	0.40	0.55	0.02	0.70	0.28	0.18	0.59	0.45
2	0.59	0.23	0.83	0.11	0.66	0.32	0.35	0.77	0.22
3	0.75	0.16	0.65	0.03	0.56	0.66	0.26	0.41	0.16
4	0.67	0.56	0.88	0.32	0.95	0.44	0.56	0.52	0.76
5	0.75	0.66	0.79	0.22	0.88	0.54	0.44	0.02	0.31
6	0.55	0.65	0.69	0.05	0.67	0.23	0.33	0.53	0.56
7	0.76	0.45	0.83	0.31	0.77	0.66	0.50	0.59	0.24
8	0.49	0.36	0.65	0.00	0.61	0.60	0.41	0.33	0.09
9	0.88	0.54	0.90	0.13	0.83	0.34	0.09	0.57	0.44
10	0.85	0.37	0.55	0.22	0.55	0.50	0.55	0.22	0.56
11	0.40	0.50	0.67	0.19	0.67	0.33	0.14	0.43	0.58
12	0.96	0.15	0.66	0.24	0.77	0.11	0.32	0.53	0.30
13	0.58	0.40	0.56	0.02	0.87	0.38	0.43	0.09	0.04
14	0.74	0.46	0.89	0.32	0.56	0.23	0.65	0.66	0.60
15	0.67	0.31	0.94	0.09	0.65	0.76	0.44	0.38	0.46
16	0.90	0.26	0.95	0.04	0.56	0.45	0.34	0.43	0.21
17	0.87	0.41	0.87	0.06	0.43	0.33	0.61	0.28	0.32
18	0.77	0.38	0.78	0.00	0.54	0.64	0.55	0.39	0.50
19	0.55	0.04	0.69	0.27	0.56	0.54	0.34	0.57	0.07
20	0.96	0.19	0.81	0.01	0.41	0.45	0.54	0.55	0.25
21	0.90	0.47	0.76	0.51	0.86	0.27	0.42	0.40	0.60
22	0.88	0.57	0.78	0.20	0.65	0.71	0.17	0.32	0.55
23	0.77	0.36	0.80	0.06	0.78	0.55	0.67	0.07	0.40
24	0.95	0.50	0.87	0.17	0.80	0.51	0.77	0.04	0.30
25	0.78	0.45	0.67	0.30	0.75	0.21	0.54	0.68	0.45
26	0.83	0.24	0.57	0.32	0.55	0,30	0.44	0.11	0.70

TABLE 3: TYPES OF ASSEMMENT TASKS AND THEIR RELIABILITY ESTIMATES

TYPE OF ASSESSMENT TASK	ESTIMATED RELIABILITY	INTERPRETATION AND POSSIBLE EXPLANATION
Timed assessment tasks: quizzes, tests, exams	Moderate → High 0.43 →0.86	These assessment tasks evoked similar responses (performance) across groups depending on students' relative rankings halfway through the course. This could be because of anxiety or differences in study strategies of high achievers, average achievers, and low achievers.
Individual, class assessment tasks: presentations, demonstrations, reporting, verbal summaries	Moderate → High 0.44 → 0.86	It seems these assessment tasks were consistently demanding to the students - to the same or similar degree depending on student's ranking halfway through the course.
Individual, take home assessment tasks: research summaries, research papers, class observations, essays, case study reports, portfolios, research proposals	Moderate 0.37 → 0.65	These tasks evoked relatively shifting ranking of students, compared to their rank halfway through the course. Students slightly shifted in ranks probably depending on the extent and quality of help they got from friends and/or family members. Secondly students may have given different degrees of focused attention to different assessment tasks.
Group, class assessment tasks: poster sessions, presentations	Moderate 0.35 → 0.61	This type of tasks reflected student's slight shifting in their respective rankings halfway through the course. Collaborating in group work might have helped weaker students to do better. Also, high achieving students might have performed lower due to group psychology.
Group, Take-home assessment tasks: long term projects, short	Low → Moderate 0.02 → 0.39	These assessment tasks seem to have been demanding to the same extents, the student's standing (ranking) halfway through the course

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC®

TM032572

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Evaluating Validity + Reliability of Classroom Assessments Using Secondary Data

Author(s): Selina Mushi, Ph.D.

Corporate Source: Northeastern Illinois University

Publication Date: April 2001

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Level 1

☒

Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.

The sample sticker shown below will be affixed to all Level 2A documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

Level 2A

□

Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only

The sample sticker shown below will be affixed to all Level 2B documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

Level 2B

□

Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, → please

Signature:

Printed Name/Position/Title:

Organization/Address: 5500 North St. Louis
Chicago IL 60625

Telephone: (773) 442-5382

FAX:

E-Mail Address: S-Mushi@neiu.edu

Date: April 13, 2001

(over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:

Address:

Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:

Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

University of Maryland
ERIC Clearinghouse on Assessment and Evaluation
1129 Shriver Laboratory
College Park, MD 20742
Attn: Acquisitions

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
1100 West Street, 2nd Floor
Laurel, Maryland 20707-3598

Telephone: 301-497-4080
Toll Free: 800-799-3742
FAX: 301-953-0263
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com