Connie Lindsey
ED 6391
June 13, 2005

Herman, J.L., Gearhart, M., & Baker, E.L. (1993). Assessing writing portfolios: Issues in the validity and meaning of scores. *Educational Assessment*, 1(3), 201 – 224.

Article Summary

Purpose: The researchers initiated this quantitative study to find out if portfolios can be scored reliably and whether those scores represent meaningful indicators of student achievement. These are important questions, according to the researchers, because many school districts and even states have adopted portfolio assessments over standardized writing assessments under the belief that portfolios improve the quality and usefulness of large-scale assessment, in spite of the fact that portfolios' measurement quality has not been established. It is, in the researchers' words "largely uncharted territory."

Method: The researchers chose an elementary school that had begun a portfolio project in 1989 in response to their participation in the national Apple Classrooms of Tomorrow. The participants of the study were three elementary school teachers and their first, third, and fourth grade students. The data covered the first year of that portfolio project and included each student's working portfolio that was originally to have contained monthly examples of narratives, as well as a standard narrative prompt that students were given 30 – 40 minutes to write on in the late spring. All teachers retained six portfolios from their classes, including two high-ability students, two medium-ability, and two low-ability students. In addition, the third-grade teacher submitted all portfolios from her class. To form two data sets, only the final drafts of narratives and summaries were used, along with the responses to the standardized prompt. Each of these samples was coded with student identification numbers and the date of the assignment; students' names and grade levels were removed. Three raters, all teachers experienced in using the analytic rubric for the district's third- and fifth-grade narrative assessments, scored the individual pieces and then assigned a competence score to each portfolio.

Results: Analyses indicated satisfactory levels of agreement for all rating contexts. Agreement was highest for the Standard Writing Assessment and the Portfolio collections as a whole (.97 to 1.00), while the lowest agreement was for Classroom Summaries. Ratings given the Standard Writing Assessments by a second set of raters in the summer showed .97 agreement between the two sets of raters. These results demonstrated that it is possible to score portfolios consistently.

The study raised serious questions, though, about the meaningfulness and generalizability of the results obtained from portfolios. There were substantial differences in students' performance among the standardized writing assessment, individual samples of classroom work, and on the portfolio collections as a whole. The portfolio competence score was consistently higher than the aggregates of the individual scored pieces, leading the researchers to believe that raters may have looked for capability and not typical

performance in scoring the portfolio as a whole. This finding raised the question of what raters are really rating: capability, best performance, or typical performance. These issues were not addressed in the rater training or on the rubric.

The researchers found several confounding variables in their study. One fundamental issue was how much help students received on work included in the portfolio. Several factors confounded the judgment of student progress: the inconsistent contents of the portfolio and the variation of task difficulty and students' familiarity with the task confounded judgments of student progress. The portfolios did not contain the monthly assignments agreed upon that would have provided raters with repeated samples of performance over time. It was also impossible to tell if a higher quality essay was the result of student growth, an easy assignment, or lots of help with editing and revising. Researchers concluded that portfolio assessment raises serious issues that need researching to assure sound assessments.

Evaluation of Research Study

Introduction: The research topic area, portfolio assessment, is clearly stated in both the first paragraph and the abstract. The problem of portfolio assessment's reliability as indicators of student achievement is clearly stated also. The literature review establishes that the term portfolio has a variety of meanings and that they differ across projects, both in purpose and who is expected to assess them. In spite of these varieties, proponents of portfolios have claimed that they can be used for numerous purposes, including large-scale assessment. The review points out, though, that assessments require consistency across raters, similar tasks, and tasks that may vary in content but still represent the same domain (all issues brought out in the research study). The review concludes by enumerating the ways that portfolio assessment departs from direct writing assessments, an assessment that has been shown to provide needed consistency. From this point, the review leads to the statement of the research questions concerning portfolio scorability, the efficiency of scoring, and the validity of scores. The research questions clearly arise out of the differences seen between standard writing assessment and portfolio assessment, and each question is clearly stated.

Methods: The participants are described only as first, third, and fourth grade elementary school students who are in the three teachers' classrooms who participated in the study. Apparently, convenience sampling was used, as the researchers found a school that was ready to initiate a pilot portfolio project with the intention of eventually including the entire school. Non-random quota sampling was also used, as teachers were asked to identify and keep two portfolios from each of three ability groups. In addition, the participant pool was broadened to include the entire class of third graders, as this teacher was identified as being most fully involved with the project. To avoid biasing the raters, writing samples were scrambled so that all grade levels were mixed.

Researchers chose rubrics as their research instruments. Rubrics seem to be a valid choice, as they have been shown to provide consistent scores on writing assignments. The selected rubric provided both holistic and analytic scores, was based on extensive

research, and had been used by a local school district for several years. Raters' reliability with the rubric was established prior to scoring with a training set of 20 samples.

Researchers were unable to control several variables, including students' prior experience with writing tasks, the amount of help received on classroom work, and missing student samples. These variables were clearly identified as confounding the assessment of student progress over time, one of the study's original goals. The issue of consent was not discussed in the report, but the participants appeared to be treated ethically, as students' names and grade levels were replaced with ID numbers.

Results: In analyzing data, a Pearson correlation coefficient was determined for rater agreement. Descriptive statistics for examining consistency of scores across contexts were reported in a table that demonstrated general evidence of the sensitivity and validity of the measure for assessing developing writing competence. In Grade 3, the analyses also provided evidence of the sensitivity of the assessments to classroom instructional emphases.

Four analyses were used to examine the comparability of student performance across the different assessment contexts: repeated measures comparisons of students' scores on the different assessments, correlations of scores among types of assessments, and two different cross-tabulations. The results of these analyses were reported in numerous tables that were clearly labeled and contained explanatory notes.

As stated earlier, raters were unable to accurately measure students' skill development over time due to confounding variables already mentioned. Therefore, no analysis of data was able to be made for that research question.

The researchers are careful to point out that the results do not indicate that results of portfolio assessment can be generalized because of the discrepancy between competency scores on portfolio collections and standardized writing assessments.

Discussion: The article contains a clear explanation of the study's results, that it is possible to score portfolios consistently, but that portfolio scores were not consistent with the aggregates of the individual scored pieces. Raters consistently gave higher scores to the whole portfolio than an aggregate of the individual scored pieces yielded. These results inform the topic of portfolio assessment, as little scientific research of its validity and reliability had been done prior to this study.

The real world of the classroom with its distractions and competing requirements limited what researchers were able to discover about portfolios' ability to reveal student growth over time reliably, as teachers were unable to follow through on their agreement to assign and collect monthly narrative assignments. This issue caused researchers to suggest that substantial monitoring and follow-up would be necessary to ensure the specified content was present in portfolios used for large-scale assessment.

Another direction for future research discussed is the need to clearly establish what raters are looking for and the need for portfolio designs to coordinate with the established scoring intent. A final implication of the study's results discussed in the article brought up the caution that the need for portfolio contents to conform to given designs in order to ensure quality assessment could hamper the very flexibility and teacher empowerment that portfolios were originally intended to encourage.

Final Words: Overall, this research study seemed to be thoughtful and well planned, a good model for studying writing assessment. The literature review reveals extensive reading in the area of both portfolios and valid large-scale assessment of writing. Although the study was non-experimental, it is significant that it was not attempted under laboratory-type conditions, as the theory being tested, whether portfolios can be reliable and valid measures of student achievement, is one that must be applied in the real world of classroom instruction. The study's results, although not generalizable, raised important questions about what competency should be assessed when rating portfolios and the need to align the portfolio's contents with the assessment intent. The most important issue raised though, may be whether using portfolios as large-scale assessment is a wise choice, even if they can be found to be reliable, valid assessments.

The very issues that confounded raters' ability to assess student growth over time are ones that teachers will probably always face: intrusions upon instructional time that disrupt plans, students with varying levels of skills, and disparity of help given to students during the editing and revising process.

Apart from the challenge presented by limited understanding of statistical analysis, the article is clearly written and thorough. Results are reported in tables for clarity, in addition to being explained in the text. The rubrics used are reproduced in the article so that the reader can see the rating criteria. It seems very likely that one could reconstruct this study with the information provided, a fact which increases the credibility of the study.