MICHIGAN READING JOURNAL

Michigan Reading Journal

Volume 34 | Issue 4 Article 5

July 2002

Is the MEAP Writing Test Reliable? A Case Study

Stephen A. Anderson

STEPHEN A. ANDERSON

Is the MEAP Writing Test Reliable? A Case Study

Stephen Anderson is the principal of Amerman Elementary School in Northville, Michigan. He is a member of Michigan Reading Association and the International Reading Association.

Although many practitioners applaud the state for stretching our students with "authentic assessments," there is also concern about the statistics and procedures used with many of these tests. A 4-point rubric raises questions about precision. How precise is a ruler with only 4 lines on it? Many of us have heard rumors of how judges are trained to grade papers based solely on "transition words" rather than content or creativity. Many of us question the validity of this test since it samples only one type of writing. Yet the state uses these "scores" politically to rank school districts and, in some cases, accuse districts of "cheating." Given the "high stakes" nature of the test, shouldn't the state be held to high standards of testing?

Because of these concerns, I have been searching for statistics that would justify the MEAP's reliability and validity. One common way to establish reliability of a test, such as the writing test that uses judges and rating scales, is to develop an inter-rater reliability correlation comparing the judgments, or scores, of each judge to see if their observations are similar. However, the state has never published such a statistic. Instead, it make its case by showing the number of cases where two scorers disagree by *more than one point* (Schram, 1999). The percentage of cases presented is very low. But remember, they are showing only the number of cases where there is a difference of two or more points. If you are only using a 4-point rubric, could a difference of one point (25 percent) make a difference? The state does not publish the number or percentage of cases where the disagreement is one point.

Kerlinger (1986) points out some of the inherent weaknesses with rating scales:

> The intrinsic defect of rating scales is their proneness to constant or biases error. This is not new to us, of course. We met this problem when considering response set. With ratings, however, it is particularly threatening to validity. Constant rating error takes several forms, the most pervasive of which is the famous halo effect. (p. 495).

Hittleman and Simon (1997) set five standards for the reliability and validity of authentic assessments. The most relevant standard to this case study is standard number 5: "The method of scoring students' knowledge, products, and performances should be clear, and there should be criteria for determining appropriate outcomes and *consistency* in the application of those criteria. (p. 162, emphasis added). What follows is a case study of the Northville Public Schools' data for the 2000 MEAP Writing test.

Case Study

One of the ways that the state now tries to resolve the problem of disagreement between judges is to average the scores of two judges. In other words, if the first judge gives a student paper a "2" and the second judge gives the same student paper a "3," the student is awarded a "2.5." Since these averaged scores are now

presented in the data presented to districts, one can reconstruct the number of cases where two judges disagree. This assumes that whole number scores represent agreement, although the number of cases where two judges disagree by two points is unknown.

Table 1(facing page) shows the frequency of scores for the MEAP 2000 Writing Test for all five elementary schools in the Northville Public Schools. Once again, if we assume that the scores of .5, 1.5, 2.5, and 3.5 indicate the disagreement of judges, one will note that the range of disagreement was from 22% to 38% of tests depending on the school. There were no scores below 1.5 in the Northville Public Schools data. The total disagreement for the district was 31% (n=139) of all 431 writing tests taken.

Using the assumptions about the interpretation of each writing test score, each of the two judges scores for each writing test was reconstructed. Reconstructing the scores for all 431 tests resulted in the frequencies shown in Table 2 (facing page).

A statistical analysis of this data using SPSS resulted in a Pearson Correlation of .606. Using Hinkle, Wiersma, and Jurs' (1988) "rule of thumb for interpreting the size of a correlation," this inter-rater reliability coefficient would be considered "moderate." The coefficient of determination (r^2) is .36. This can be interpreted to mean that only 36 percent of the variance in Judge 1's scores are related to, or associated with, the variance in Judge 2's scores.

Discussion

The state is using these scores to make determinations about accreditation, penalties, and relative achievement between districts. In addition, the scores are released to the media to rank schools. These are extremely high stakes, and this case study raises serious questions about the reliability of the test instrument. Are the scores consistent? The data indicate irregularities when disagreements can vary from 22 percent to 38 percent of all tests in a building and almost one out of every three (31 percent) in the district. One questions the probability of the scores when none of the district's identified gifted students (n=47) received a "4."

Given the high stakes nature of these tests, can we be satisfied with a "moderate" inter-rater reliability coefficient? Remember that this analysis is based upon the assumption that there were no disagreements of greater than 1 point since this data is not presented by the state. If such cases existed, the correlation would have been lower. Aiken (1997) in answering the question of how large a reliability coefficient should be answered:

> If interest is limited to differentiating between groups of people, then a coefficient of .70 may be sufficient. But if we want to differentiate between or within individuals, then a coefficient of at least .85 is probably necessary. (p. 158)

Is the MEAP writing test reliable? In this case study the state fails both in terms of consistency and the size of the inter-rater reliability coefficient.

Bibliography

Aiken, L. (1997). Questions and inventories: Surveying opinions and assessing personality. New York: John Wiley & Sons, Inc.

Hinkle, D.; Wiersma, W.; and Jurs, S.(1988). *Applied statistics for the behavioral sciences*. Boston: Houghton Mifflin Co.

Hittleman, D. and Simon, A. (1997). *Interpreting educational research*. Upper Saddle River, NJ: Prentice-Hall, Inc.

Kerlinger, F. (1986). *Foundations of behavioral research*. New York: CBS College Publishing.

Schram, C. (1999, Nov.). Measurement 101. *MEAP Update*, *12*, (2), 5.

Table 1
School and District Frequency Distributions

School	Scores												% Disagree	
	1.5		2.0		2.5		3.0		3.5		4.0			
	N	%	N	%	N	%	N	%	N	%	N	%	N	%
Amerman	2	2	41	37	31	28	31	28	6	5	0	0	**39**	**35**
Moraine	0	0	6	11	19	33	29	51	3	5	0	0	**22**	**38**
Silver Springs	3	3	25	28	23	26	34	38	4	4	0	0	**30**	**33**
Thornton Creek	0	0	29	36	25	31	25	31	2	2	0	0	**27**	**33**
Winchester	1	1	16	17	18	19	54	58	2	2	2	2	**21**	**22**
District	6	1	117	27	116	26	173	40	17	4	2	>1	**139**	**31**

Table 2
Frequencies of Reconstructed Scores

	Score	f	%	C%
Judge 1	1	6	1.4	1.4
	2	233	54.1	55.5
	3	190	44.1	99.5
	4	2	.5	100.0
	Total	431	100.0	
Judge 2	1	0	0	0
	2	123	28.5	28.5
	3	289	67.1	95.6
	4	19	4.4	100.0
	Total	431	100.0	

Call for Manuscripts

Summer, 2003: Creating Professional Learning Communities

(Manuscripts must be received by January 1, 2003. Electronic submissions are encouraged.)

The knowledge base teachers draw upon when selecting instructional tasks and materials has an important impact on the quality of student learning that occurs in their classrooms. According to the *National Reading Panel: Teaching Children to Read*, there is a growing body of research describing the correlation between aspects of teacher preparation and the quality of teaching and student outcomes. Ongoing professional development is an essential component of a school or district's early literacy plan. Teachers need the opportunity to expand their knowledge base and to increase skill in instruction if they are going to meet the ever-changing educational needs of their students. Teachers also need to belong to a learning community in which they share their knowledge, skill, and insight with other teachers as they continuously strive to provide effective instruction for each student in their building.

The goal of this issue is provide information about the context, content, and process of effective professional development. The focus of the articles in this issue is sustained professional development with the goal of continuous professional growth leading to increased student achievement. Journal editors seek manuscripts for this issue describing the establishment and maintenance of learning communities that provide opportunities for teachers to share their knowledge, skill, and insight with colleagues as they continuously strive to provide effective instruction for each student in their building. Contributors are invited to send articles describing successful professional development programs in their districts or schools, processes used to establish learning communities within and across districts and schools, and teacher research and inquiry projects that have added to the knowledge base supporting initiatives in their schools.

- Manuscripts should not exceed 2,500-3,000 words.
- Author's name, mailing address, telephone number, FAX number, e-mail address, and professional affiliation should be on a separate cover page. The author's name should not appear in the manuscript.
- Three members of the editorial review board will review all manuscripts.
- Manuscripts must be received by January 1, 2003. Decisions will be reached within four months for this issue.
- If a manuscript is accepted for publication, its author must provide a computer disk copy of the manuscript, preferably in MS Word.
- Charts, graphs, drawings, and high quality photographs pertaining to article topics will be appreciated. Photographs from a digital camera can be submitted digitally.
- Send six copies of the manuscript and two self-addressed, stamped envelopes to:

  Kathleen Clark, The *Michigan Reading Journal*
  Oakland University Department of Reading & Language Arts
  Rochester, MI 48309-4494