DOCUMENT RESUME

ED 344 931 TM 018 275

AUTHOR Tibbetts, Katherine A.; And Others
TITLE Development of a Criterion-Referenced, Performance-Based Assessment of Reading Comprehension in a Whole Literacy Program.
INSTITUTION Kamehameha Schools/Bernice Pauahi Bishop Estate, Honolulu, HI.
SPONS AGENCY Department of Education, Washington, DC.
PUB DATE Apr 92
CONTRACT S208A900001
NOTE 17p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 20-24, 1992).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Constructed Response; *Criterion Referenced Tests; Educational Assessment; Formative Evaluation; Grade 3; Inservice Teacher Education; *Literacy; Outcomes of Education; Portfolios (Background Materials); Primary Education; *Reading Comprehension; *Student Evaluation; Summative Evaluation; *Test Construction; *Whole Language Approach
IDENTIFIERS Kamehameha Early Education Program; Kamehameha Schools HI; *Performance Based Evaluation

ABSTRACT

This paper describes the development of a criterion-referenced, performance-based measure of third grade reading comprehension. The primary purpose of the assessment is to contribute unique and valid information for use in the formative evaluation of a whole literacy program. A secondary purpose is to supplement other program efforts to communicate and reinforce objectives for student performance and instructional practices. The Kamehameha Schools/Bishop Estate is a private non-profit educational institution in Hawaii. One of its largest and oldest projects is the Kamehameha Early Education Program (KEEP), which is designed to improve the literacy skills of native Hawaiian children by improving the quality of instruction they receive. KEEP hires and trains teacher consultants for public elementary schools to provide training and support. An innovative student assessment was developed to determine student outcomes supplementing a portfolio approach with a criterion-referenced test with performance-based constructed response items. Assessment development was a collaborative effort of educators, students, and evaluators that was field-tested in 1991. The prototype assessment directly taps curricular objectives in a format that is congruent with instructional practices. There is a 14-item list of references. (SLD)

Development of a Criterion-Referenced, Performance-Based Assessment of Reading Comprehension in a Whole Literacy Program

Katherine A. Tibbetts

Anita A. Peterson

Wendie C. Yumori

Kamehameha Schools/Bishop Estate

Honolulu, Hawaii

Symposium paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA, April, 1992.

Running Head: Performance-Based Assessment

The research reported here is supported in part by the *Native Hawaiian Model Curriculum Implementation Project* funded under grant number S208A900001 from the U. S. Department of Education and administered by the Kamehameha Schools/Bishop Estate.

Abstract

This paper describes the development of a criterion-referenced, performance-based measure of third grade reading comprehension. The primary purpose of the assessment is to contribute unique and valid information for use in the formative evaluation of a whole literacy program. A secondary purpose is to supplement other program efforts to communicate and reinforce objectives for student performance and instructional practices.

Setting

The Kamehameha Schools/Bishop Estate (KS/BE) is a private, non-profit educational institution in the state of Hawaii. The Early Education Division (EED) of KS/BE alone services over 6,000 children and families each year. One of the oldest and largest of the KS/BE EED programs is the Kamehameha Early Education Program (KEEP). KEEP's mission is to improve the literacy skills of native Hawaiian children by improving the quality of instruction they receive. To accomplish this, KEEP hires and trains teacher-trainer consultants who are placed at public elementary school sites to provide training and support to teachers in the implementation of KEEP. The public schools are selected from schools in the state that 1) serve a high proportion of low-achieving Hawaiian/part-Hawaiian students, and 2) are receptive to the program. KEEP is currently implemented in over 130 classrooms in eight public elementary schools in four school districts.

In the 1989-90 school year, KEEP began the process of moving from a mastery model of instruction to a whole-literacy based model. In response to these program changes, the Evaluation Department reviewed the KEEP evaluation design and instrumentation. As a result of that review, a number of changes have been proposed and implemented, including the development of the innovative student assessment reported in this paper.

Needs

Key to the revised, formative KEEP evaluation design are student outcomes. To what extent is the program successful in meeting its own objectives for student performance? What teacher, student, school, and community characteristics are related to achievement of intended outcomes (and to what extent can educational programs reduce any negative effects)? What are the long-term effects of KEEP participation on student achievement? Is there a positive relationship between meeting KEEP objectives and performance on the norm-referenced tests of reading? To answer these questions, a valid and reliable indicator of student performance relative to intended outcomes is needed.

Prior to the move to a whole-literacy model, KEEP student outcomes were measured with standardized tests such as the Stanford Achievement Test (SAT) and with a series of skill-oriented, multiple-choice, criterion referenced tests (KROS). As part of the transition to whole literacy the curriculum developers (the KEEP Project Team) replaced KROS with an extensive classroom assessment package using a portfolio approach (Au and Blake, 1992).

Review of the SAT and the portfolio system revealed a gap in the type of information critical to answering the evaluation questions described above. As suggested by others, (Frederiksen and Collins, 1989; Hambleton, 1991; Hiebert and Calfee, in press; Nitko, 1989; Shepard, in Kirst, 1991), we have found the norm-referenced tests to have limited use in program evaluation. When reading was conceived of and taught as a series of discrete skills that could be mastered and assessed in isolation, there was a reasonable match between curriculum and the standardized multiple-choice achievement tests. (See Resnick and Resnick, in press, for a discussion of this issue.) However, the match between the KEEP whole literacy curriculum and those same tests is very poor and valid inferences can not be made from SAT scores about student achievement of specific KEEP curricular objectives. For example, the SAT assesses some skills (such as phonemic awareness) in a

manner antithetical to whole literacy instructional practices. The SAT administration would be the first time this task would be presented in isolation to KEEP students unless time was taken from regular instruction to drill this item type. In the area of comprehension, the SAT passages used are too short to support the type of questioning relevant to curricular objectives, i.e. discriminating between more and less important traits of main characters, identifying a problem and explaining why a situation is problematic for the character, and discerning author's message.

The KEEP portfolio assessment is designed, among other things, to communicate the curriculum and consequently has high face validity. However, lack of consistency in teachers' understanding and application of standards is a source of unreliability and is problematic for program evaluation. Variation in portfolio ratings also occurs due to the different expectations individual teachers have for student performance based on their personal beliefs. These expectations are mediated in part by the variation in general achievement levels between schools. The situation is exacerbated by the move to more holistic and, therefore, less clearly specifiable outcome objectives. The Project Team and teacher-trainer consultants are actively addressing these issues and striving for ever improved consistency (Au and Blake, 1992).

Based on the above assessment of existing data sources relative to program needs, the Evaluation Department proposed the development of a third indicator of literacy progress. The assessment results will be more valid indicators of program effects and more reliable than the portfolio. The practice of "teaching to the test" is pervasive and well documented (Shepard in Kirst, 1992; Smith, 1991). To ensure the test did not detract from program dissemination efforts or the students' educational experiences, it was important to develop "a test worth teaching to." Consequently, a secondary function of the assessments is to communicate and reinforce selected KEEP objectives for both student performance and instructional practices (Brauer and Hiebert, 1991).

Using criterion-referenced test development practices to develop an assessment which includes constructed response item formats furthers both functions. Criterion-referenced test development procedures would help ensure the assessment results were relevant to curricular objectives. Using performance-based (constructed response) item formats allows students to be assessed on the same types of tasks that are used in daily instruction and, again, helps to ensure the results are relevant to the program while reinforcing the KEEP instructional model. (See Osterlind (1988) for a discussion of the use of criterion-referenced assessment in program evaluation.)

The test results are intended for use by teachers, teacher-trainers, curriculum developers, policy makers/administrators. It is expected that teachers will find that the tests provide a useful piece of information for their assessment of the progress of individual students. It is also expected that teachers will find the class profiles helpful in identifying areas of strength and weakness for the purpose of making classroom-level instructional decisions. Teacher-trainers can use the results to identify areas to emphasize in their work with teachers. Classroom-level and school-level results will be used in conjunction with other data to evaluate program effectiveness.

Methods

The assessment development has been a collaborative effort of students, teachers, teacher-trainers, curriculum developers, and program evaluators. While primary responsibility for the project has rested with Evaluation Department staff, the expertise of the whole group has been essential to this project.

The first step of the process was to increase in-house expertise in the development of criterion-referenced assessments. Dr. Ronald Hambleton was contracted to provide a five-day workshop in this area and to provide ongoing consultation through the first 18 months of the development process.

The next step in the assessment development entailed discussion about the specific aspects of the instruments to be developed. Key decisions were made concerning 1) the first aspect of literacy to assess (reading comprehension), 2) the first grade level to develop a measure (grade 3), 3) evidence of comprehension (domains to include or exclude), 4) whether standards would reflect time of year, student reading level or end-of-year expectations (end-of-year), 5) assessment format (paper and pencil with a mix of constructed response and multiple choice items), 6) the number of forms (four), and 7) procedures (independent work to be completed in two sittings). However, these decisions were easy compared to the tasks of identifying a pool of texts meeting assessment criteria and drafting items. All of these decisions have been revisited a number of times in the development process as new questions have arisen and experiences have caused us to reconsider earlier assumptions.

The first draft assessment was developed with the assistance of curriculum developers and teacher trainers who were hired to work an extra week during the summer with evaluators on this project. These relatively "assessment naive" staff were wonderfully challenging of measurement assumptions and initial assessment blueprints and domain specifications. In return, the "instructionally naive" evaluators offered close questioning of grade level benchmarks and preferred instructional practices.

The first draft assessment was pilot-tested in July, 1990 with two summer school classes. The objectives of the pilot-test were to assess the suitability of the proposed texts, the testing instructions and formats, and the clarity and appropriateness of individual items. Teachers participating in the pilot-test were asked to evaluate the text, directions, and test items. The time students required to read the texts and respond to questions was recorded. Students' questions were noted and used along with their written responses and teachers' observations to improve the test procedures, instructions, and items. Following the pilot-test, all aspects of the assessments were

revised, going back to the test blueprint and domain specifications, although no substantiative changes were necessary at the conceptual level.

In Fall 1990, the revised assessments were field-tested with fourth grade students (the closest approximation available to exiting third grade students). The objectives for the field-test were the same as those for the pilot-test with the addition of standard setting. Actual student responses were compared to end-of-grade 3 benchmarks from the KEEP curriculum guide (Au, etc. 1991) to refine the scoring schema and identify examples of different levels of performance in reading comprehension. Curriculum developers and evaluators worked together in the initial development of scoring schema. Preliminary discussions about scoring centered around holistic impressions of quality, "this is a better answer than that." Next, response attributes were identified as raters justified how they discriminated between responses. These bases for discrimination were checked against grade level benchmarks and codified into the scoring schema. Traditional item analysis techniques were used to examine the functioning of both constructed response and multiple choice items and to revise prompts and distractors where indicated. The blueprint and domain specifications were also revised to reflect refinements in the bases for discerning the "quality" of student responses. Dr. Hambleton returned to assist in this process.

In January, 1991, the revised assessment was field-tested with third grade KEEP students. At the same time we collected information on teacher ratings of student reading levels and SAT reading comprehension scores to use as indicators of concurrent validity. The same review and revision process described in the case of the Fall field-test was used to further refine all aspects of the assessment.

What was expected to be a final field-test was conducted in May, 1991. Again the review and revision process was completed with the assistance of off-contract teachers and curriculum developers rehired for this purpose.

When planning for implementation of the assessments in Fall 1991, program staff voiced new concerns about the comparability of text selections. Two of the texts had somewhat implicit story elements while the story elements in the other texts were largely explicit. Students had more difficulty with the assessment tasks for the implicit stories. The difference in scores between the explicit and implicit stories was statistically significant on some domains. Therefore, new texts were selected to replace the two more implicit stories. The new texts have been reviewed by program staff and "child-tested." Draft questions are currently being field-tested in preparation for the May administration.

Concerns about testing procedures have generally pertained to congruence between assessment and instructional practices. Variations in testing procedures have been discussed extensively with program staff and practitioners. Teachers will now have the option of keeping the texts used in the assessment in their classrooms following the formal assessment and using the reading/assessment assignment as an instructional opportunity. Initially, this was thought to be too intrusive, but it is now considered desirable because it "makes the assessment more like instruction."

Results

The Assessment

The current version of the assessment uses both constructed response and multiple-choice item formats to assess reading comprehension. The assessment begins with activities designed to assist students in tapping their experiences and prior knowledge in preparation for reading the text. The Experience component of the Experience-Text-Relationship (ETR) model of comprehension instruction that has been a key cornerstone of KEEP (Au, 1979). The students then read the text, a complete story from children's literature that requires 15 to 20 minutes for most children to read.

After reading, the students are asked to write about a connection between the text and their lives, and respond to the more traditional comprehension items. There are six domains represented in the assessment: personal response, character description, problem identification, problem resolution, author's message (and application/connection), and word meaning from context. The personal response and application/connection domains are assessed with two constructed response items. The character description, problem identification, and problem resolution domains are assessed by one constructed response and one multiple choice item each. For problem identification and problem resolution domains, the item prompts instruct students to explain or justify their responses. Deriving word meaning from context is assessed by two constructed response and two multiple choice items.

Items are scored based on the depth of comprehension reflected by the responses. Scores range from zero points for incorrect responses (no reasonable basis for the response in the story) to the maximum points (three or five) for responses that go beyond concrete elements or story-specific details. Item scores of two, for a three-point item, or three, for a five-point item, mean the responses demonstrate concrete or literal comprehension.

The multiple choice items are written with one incorrect but credible option and three correct options that reflect increasing levels of understanding corresponding to the scoring schema. The correct responses are worth one, two, or three points. Whenever possible, the multiple choice options have been selected from actual student responses to open-ended questions.

The use of related constructed response and multiple choice items makes it possible to identify children who can recognize a "higher level" response than they can produce on their own. More detailed information on the items are included in the domain specifications, available from the authors.

Application

As stated above, the goal of these test development efforts is to produce data to enhance the utility of formative program evaluation. Linn, Baker, and Dunbar's (1991) article, "Complex, Performance-Based Assessment: Expectations and Validation Criteria," provides a cogent discussion of expanded definitions of validity that are particularly relevant to evaluative use of assessment data. Therefore, we have used the criteria for evaluation assessments proposed in their article to provide the context for discussing the validity of our instrument.

Consequences: "intended and unintended effects of assessments on the way teachers and students spend their time and think about the goals of education." (Linn, Baker, and Dunbar, 1991, p.17). We have not had sufficient experience with the assessment to respond with assurance to concerns about unintended effects. However, this point has received substantial attention throughout the assessment development process. The intended effects on the time allocation and goals of the teachers and students are clear. The assessment is tied very tightly to the curriculum in both content and form. If the curriculum is "worth teaching," and we believe it is, then the assessment is "worth teaching to."

Unfortunately and unavoidably, the assessment is of narrower scope than the curriculum. It does not encompass all of reading comprehension (even of realistic fiction), much less all of literacy. There is the potential for the assessment, if given disproportionate weight, to narrow teacher and student focus. To reduce the likelihood of this unintended effect, we will work to ensure results are used appropriately. In addition, we will continue to observe classroom instruction and speak with students, teachers, and curriculum developers to watch for evidence of unintended effects.

Fairness: "Gaps in performance among groups exist because of difference in familiarity, exposure, and motivation on the tasks of interest. . ." (Linn, Baker, and Dunbar, 1991, p.18) and "performance ratings reflect the examinee's true capabilities and are not a function of the

perceptions and biases of the persons evaluating the performance." (Stiggens, 1987, p. 33; cited in Linn, Baker, and Dunbar, 1991, p.18). The potential for "unfair" gaps in performance exists with this instrument, as in all others. A major threat to fairness is the degree to which the teacher does not adhere to the curriculum. Students of teachers who do not follow the curriculum guide, who have standards of performance lower than the benchmarks, or whose instructional practices do not afford students opportunity to practice responding to items like those in the assessment are at a disadvantage. Students in these classrooms may not fully understand what is being asked of them and the assessment results will not reflect their actual achievement. For program evaluators and developers, poor test results provide a means to identify teachers who may need more assistance to implement KEEP.

The appeal of the stories to students and the congruence between student backgrounds (as a group) and the background knowledge required to understand the story fully are used in story selection to help reduce effects of experience or motivation. In addition, it is believed that encouraging the teachers to "teach" the stories after the assessment will have a positive effect on student and teacher attitudes toward the assessment.

The use of qualitative standards in scoring sets this assessment apart from others we have seen and increases its utility within the context of the KEEP program. However, the qualitative scoring opens the door wider to potential bias in scoring. Avoiding bias in the scoring of student responses and the ranking of the multiple choice items is challenging. It is important to keep asking, "what is reasonable to expect from a nine-year-old child?" In developing the scoring schema, we have had the benefit of working with individuals representing a wide range of professional backgrounds, experiences, and perspectives. We have also used actual student responses wherever possible in writing the scoring schema and multiple-choice options. This grounds the assessment in the reality of meanings the children construct for themselves. We believe

that these efforts have helped to reduce bias. As more data becomes available through the actual use of the assessments, we will systematically monitor for bias by examining empirical data and listening critically for bias in our discussions of responses.

Transfer and generalizability: "Scores . . . become interesting only to the extent that they lead to valid generalizations about achievement more broadly defined" (Linn, Baker, and Dunbar, 1991, p. 18). The limited generalizability of many performance based assessments makes us cautious about over-interpreting our results. We have found satisfactory statistical correlations with both teacher ratings of reading level ($r=0.60$, $p=0.001$) and SAT reading comprehension scores ($r=0.66$, $p=0.001$). We also have a satisfactory correlation between the January and May field-tests ($r=0.61$, $p=0.001$).

A limitation of the assessment is the effect of genre. The stories selected are all realistic fiction. We aspire to develop assessments based on informational text at a future date and, ideally, to create a text set and assessment that requires integration of the two genres.

An additional constraint, at the student level, is the small number of items and the restriction to one text passage per test form. This results in a potentially less accurate rating than is available from less intensive, but more extensive assessments, such as traditional achievement or criterion-referenced tests. The results of this assessment should only be used in conjunction with other information in making inferences about individual students.

Cognitive complexity: "Judgements regarding the cognitive complexity of an assessment need to start with an analysis of the task; they also need to take into account student familiarity with the problems and the ways in which students attempt to solve them." (Linn, Baker, and Dunbar, 1991, p.19). The complexity of the test items vary by domain. Generally, the scoring schema give more credit for responses that demonstrate comprehension that goes beyond a literal understanding of the story. A number of questions require an understanding of the dynamics of relationships and

character development to answer well. The word meaning from context items generally require that the student attend to an entire paragraph at least, to construct or recognize a response that receives the maximum points.

Content quality: *"The content needs to be consistent with the best current understanding of the field and . . . reflective of what are judged to be aspects of quality that will stand the test of time. . . . the tasks selected to measure a given content domain should themselves be worthy of the time and efforts of students and raters." (Linn, Baker, and Dunbar, 1991, p.19).* The collaboration between professionals with expertise in instructional methodology, content area, and assessment contributes to the content quality of this instrument. The assessment is highly congruent with the constructivist principals underlying the whole literacy movement in general and KEEP in particular. The teachers, teacher-trainer consultants, and the curriculum developers have been very generous with their time and commitment to work with the evaluators to make this, in the words of one consultant, "the best test it can be."

Content coverage: *"breadth of coverage should not be overlooked." (Linn, Baker, and Dunbar, 1991, p.20).* Substantial thought has been given to the selection of the domains used to represent the much broader reality of reading comprehension. Program staff have had a major role in selecting domains for inclusion. We believe the domains assessed provide a satisfactory basis for making generalizations about reading comprehension. Future plans include the development of assessments using other genre.

Meaningfulness: *"students deal with meaningful problems that provide worthwhile educational experiences." (Linn, Baker, and Dunbar, 1991, p.20).* If the assessment and the curriculum it is intended to assess have larger social validity (Heath, 1991) is a question only time can answer. The KEEP curriculum attempts improve the literacy of low-achieving minority students by bringing to their classrooms the best current knowledge and theory about literacy and

instruction. In using the KEEP instructional practices and outcome objectives as a basis for the test we have made the assumption that these instructional practices are, in fact, meaningful.

Cost and efficiency: "ways must be found to keep the costs at acceptable levels." (Linn, Baker, and Dunbar, 1991, p.20). Developing this assessment has required a substantial investment of time and money. Costs to date include two years of staff time, money to purchase the rights to reproduce texts or buy them, money for consultants to increase our expertise to an acceptable level, and money for scoring. This assessment will continue to be more expensive to use than the strictly multiple choice tests used in the past.

We intend to train teacher-trainer consultants and teachers in scoring their students' responses. This training will provide them with more direct awareness of student responses, a benefit to them. It also reduces the scoring costs incurred by the Evaluation department. If the expense is shared across the organization, it becomes more manageable.

Evaluation would continue to collect and rescore a sample of tests. These rescored tests would be used as program evaluation data and to identify sites or teachers whose scoring is divergent from the schema. This would enable us to retrain school staff or to modify the schema if they have made significant improvements.

Discussion and Conclusions:

This extensive, collaborative process has resulted in the development of a prototype assessment of reading comprehension that directly taps curricular objectives in a format that is congruent with instructional practices. We believe it will contribute important and unique information that we need to respond to the formative evaluation questions identified at the beginning of this paper. It will also help to communicate and reinforce the importance of the curriculum benchmarks. However, a number of issues related to validity remain unanswered or

only partially addressed. We need to continue to attend to these issues; we must not assume that our good intentions are sufficient.

The Evaluation Department could not have created an instrument of this quality working in isolation. The availability and commitment of curriculum and instructional staff to working on this project have been critical to its success.

It has been a lengthy, expensive process. But, the process itself has been valuable. Evaluation has a deeper understanding of the whole literacy program and the program has had to define more clearly its standards and practices.

References

Au, K. H. (1979). Using the experience-text relationship method with minority children. *The Reading Teacher, 1*, 677-679.

Au, K. H. (1990). From reading to literacy: changes to the KEEP curriculum. *The Kamehameha Journal of Education, 1(1)*, 1-8.

Au, K. H., & Blake, K. M. (1992, April). *Development and implementation of a whole literacy-based curriculum framework and portfolio assessment system.* Paper presented at the meeting of the American Educational Research Association, San Francisco, CA.

Brauer, K. J., & Hiebert, E. H. (1991, December). *A Critical Review of Newer State and National Assessments: Do They Measure Up as Classroom Assessments?* Paper presented at the National Reading Conference, Palm Springs, CA.

Frederiksen, J. R., & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher, 18(9)*, 27-32.

Hambleton, R. K., & Murphy, E. (1991). Changes in Educational Testing Practices. *Kamehameha Journal of Education. 2(2)*, 17-23.

Heath, R. W. (1991). Authentically what? *Kamehameha Journal of Education. 2(2)*, 27-36.

Hiebert, E. H., & Calfee, R. C. (in press). Assessment of literacy: from standardized tests to performances and portfolios. In S. J. Samuels & A. Farstrup (Eds.). *What Research Says About Reading Instruction (2nd Ed.).* Newark, Delaware: International Reading Association.

Kirst, M. (1991). Interview on assessment issues with Lorrie Shepard. *Educational Researcher, 20(2)*, 24-27.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: expectations and validation criteria. *Educational Researcher, 20(8)*, 15-21.

Nitko, A. J. (1989) Designing tests that are integrated with instruction. In R. L. Linn (Ed.) *Educational Measurement, Third Edition,* New York: Macmillan Publishing Company.

Osterlind, S. J. (1988). Using CRTS in program curriculum evaluation. *Educational Measurement Issues and Practice 7(3)*, 23-30.

Resnick, L. B., & Resnick, D. P. (in press). Assessing the thinking curriculum: new tools for educational reform. In B. R. Grifford & M. C. O'Connor (Eds.) *Future Assessments: Changing Views of Aptitude, Achievement, and Instruction.* Boston, Kluwer Academic Publishers.

Smith, M. L. (1991) Put to the test: the effects of external testing on teachers. *Educational Researcher, 20(5)*, 8-11.