Developing a Valid and Reliable Portfolio Assessment in the Primary Grades: Building on Practical Experience

Kelly S. Shapley and M. Joan Bush

Dallas Public Schools
Dallas, Texas

In response to assessment guidelines defined by the Improving America's Schools Act of 1994, a reading/language arts portfolio assessment was included in the Dallas Public Schools' Title I Local Education Agency Plan (Dallas Public Schools, 1995) as an optional student assessment for students in prekindergarten to second grade. The objective of this study was to examine the extent to which the 1995–96 portfolio assessment met appropriate technical standards for its intended uses. After 3 years of development and gradual improvement, the portfolio assessment did not provide high quality information about student achievement as hoped. The reliability of the scores was low, and the portfolio contents did not provide a valid sample of students' work. Implications for other districts are discussed related to technical quality, staff development, and cost.

Current national and state educational initiatives promote challenging academic standards for all students. In particular, the Title I legislation urges the use of alternative assessments to better measure what students know and are able to do (Improving America's Schools Act of 1994). As a result, the Dallas Public Schools elected to develop and implement a reading/language arts portfolio in the primary grades to provide an assessment closely aligned with the Texas content standards. The portfolio is a collection of student work used in the assessment of reading and writing. Work samples (e.g., writing, story retellings, informal assessments, tape-recorded readings, etc.) provide evidence of students' literacy development

over time. The portfolio provides a form of assessment that is embedded within instruction and learning and entails the sampling of a broad range of students' work. Thus, portfolios enable teachers to tap the multidimensionality of reading and writing, which involves students' cognitive processes, affective responses, and literary products.

Portfolios are only one of a proliferation of performance assessments that advocates are currently lauding for their authenticity and instructional relevance. Such assessments are seen as having the potential to positively influence teaching and learning, portray students' thinking processes, align with important outcomes, capture ongoing progress, integrate assessment and instruction, involve students in their own learning as well as provide valuable information to parents and policymakers (see, e.g., Arter & Spandel, 1992; Frechtling, 1991; Herman & Winters, 1994; Tierney, Carter, & Desai, 1991). Unfortunately, portfolios have often received educators' support based on what they appear to measure rather than on what portfolios actually measure (Mehrens, 1992). Despite educators' optimism, limited empirical evidence exists supporting the cited benefits, validity, or reliability of portfolio scores (The Evaluation Center, 1994; Gearhart & Herman, 1995; Herman & Winters, 1994, Koretz, Stecher, Klein, & McCaffrey, 1994; Koretz, Stecher, Klein, McCaffrey, & Deibert, 1993).

Evidence on the reliability and validity of portfolio scores comes primarily from writing and mathematics portfolios in the upper elementary and higher grades (e.g., Vermont's portfolio assessment, Koretz et al., 1994; Pittsburgh's district-wide writing portfolios, LeMahieu, Gitomer, & Eresh, 1995; the Kentucky Writing Portfolio, The Evaluation Center, 1995). Technical quality results for portfolios are mixed, but generally cautions have been raised about the utility of portfolios for accountability decisions. A search of the literature reveals that there is scant evidence available on the technical quality of portfolios used in the primary grades. Available information describes the use of portfolios to support classroom instruction; however, little is known about the credibility of portfolio contents or scores for assessing primary children's performance.

ENSURING THE QUALITY OF PORTFOLIO SCORES

The challenge of using portfolios is to develop an assessment that is adequately reliable and valid for both instructional and informational purposes. Technical standards need to be more rigorous for high-stakes student-, teacher-, and school-level decisions than for low-stakes use of results such as providing information about the status of the educational system (Linn, 1994). Still, for important decisions to be sound, portfolios must yield accurate, reliable, consistent, fair, and meaningful information (Herman & Winters, 1994; Reckase, 1995). Attention to technical qualities such as reliability and validity is essential when designing classroom assess-

ment systems that purport to advance instructional reform and measure students' progress toward valued content and performance standards.

Reliability

Reliability concerns the consistency or stability of scores. In the case of portfolio assessment, there must be "interrater agreement." This refers to the consistency with which two or more judges using shared standards rate the same performance. Higher levels of interrater reliability are more readily achieved when (a) the contents of the portfolio are relatively uniform—tasks are clearly specified and standardized, (b) the scoring rubrics are well defined, (c) raters thoroughly understand the performance criteria, and (d) raters are experienced and well trained. Additionally, portfolio score reliability is related to the consistency of scores across different occasions or tasks (Herman, Aschbacher, & Winters, 1992; Herman & Winters, 1994; LeMahieu et al., 1995).

The decisions to be made (high or low stakes) dictate the level of interrater agreement that is acceptable; however, in general, reliability coefficients of .70 are considered respectable for portfolio assessments (Herman et al., 1992). Interrater reliabilities from Vermont's statewide portfolio writing assessment (.28 to .57) were inadequate for school or district comparisons (Koretz et al., 1993). On the contrary, the Pittsburgh portfolio assessment scores in writing showed higher interrater agreement (.74 to .87) as a result of extensive rater training and an arbitration process (LeMahieu et al., 1995). In another study, a similar level of interrater agreement (.82) was reached by pairs of raters assigning elementary writing portfolios a single overall score (Herman, Gearhart, & Baker, 1993). Data on the reliability of portfolio scores for students in the primary grades are limited. Shapley and Pinto (1995) found that it was impossible to calculate interrater agreement for many portfolio scores because the unstandardized tasks placed in primary portfolios provided inadequate evidence for second raters to judge students' performance. On the other hand, Meisels (1997), using a Work Sampling System, reported that a checklist and summary report (including portfolio ratings) had moderately high interrater reliability.

Validity

Reliable scoring is necessary, but the central issue in judging the technical quality of an assessment is validity. However, without adequate reliability, validity is undermined. Specifically, validity evidence is needed to support inferences that people make about student performance and to determine whether scores represent some enduring or meaningful capability or a generalizable skill. Data must show

that an assessment measures what it is supposed to measure and that the interpretation of the assessment is justified. To do so, evidence must be accumulated to demonstrate that there is sufficient sampling to permit judgment. Additionally, patterns of relations between portfolio scores and other indicators of student performance must be examined to determine criterion-related validity. To establish concurrent criterion-related validity, convergent and divergent evidence must be reviewed. Portfolio scores should relate strongly to other accepted and valued measures of the same capability (convergent validity) and demonstrate weak or no relation with measures of different capability (divergent validity; Herman & Winters, 1994; Reckase, 1995).

The validity of portfolio assessment scores often remains questionable. Koretz et al. (1993) found only moderate correlations (.47 to .58) between Vermont's writing portfolio scores and other writing assessments. Gearhart, Herman, Baker, and Whittaker (1993) found virtually a zero relation between portfolio writing scores and the results of a standard writing assessment. Shapley and Pinto (1995) found low correlations (.17 to .39) between primary students' reading/language arts portfolio mastery ratings and standardized reading and language subtest scores.

PURPOSE OF THE STUDY

The Reading/Language Arts Portfolio Assessment was included in the Dallas Public Schools' Title I Local Education Agency Plan as an optional student assessment for students in prekindergarten to second grade. The main purposes of the portfolio were the ongoing documentation of students' progress in meeting state performance standards and the improvement of instruction and learning for students. The objective of this study was to examine the extent to which the 1995–96 portfolio assessment met appropriate technical standards for its intended uses. Specifically, the study addressed (a) the reliability of the portfolio assessment scores, (b) the evidence that supported the validity of inferences made from portfolio scores, and (c) the implications for further development and refinement of the portfolio assessment.

DISTRICT-WIDE PORTFOLIO ASSESSMENT

Portfolios were first used in the district when the Texas Education Agency allowed the use of alternative outcome measures for assessing the performance of Chapter 1, now Title 1, students. Previously, gain scores on standardized tests were used to assess program effectiveness. In the first 2 years of implementation, 1992–93 and 1993–94, students' portfolios were informal and unstructured and could best be described as "a classroom collection of student work." Guidelines simply asked teachers to include samples of student work and a grade-level checklist in a folder.

Students' mastery of the grade-level content standards was defined by a response of "most of the time" or "some of the time." A response of "not yet" placed the student in the nonmastery category (Boykin, Yang, & Benoit, 1993; Sheehan, Yang, Shapley, Johnson, & Thapa, 1994).

In 1994–95, a portfolio development initiative focused on making the portfolio a more valid, reliable, and equitable measure, that is, a portfolio assessment with an established purpose, criteria to determine what is put into the portfolio, by whom, and when, and criteria for assessing the work (Herman et al., 1992). Title I instructional specialists and evaluation staff members collaborated to identify and define key student outcomes, create scoring rubrics, design an ongoing record of the student's progress at specified time points, and set expectations for portfolio contents. Unique goals and scoring criteria were outlined for English proficient and limited English proficient students (Shapley & Pinto, 1995).

The 1995–96 version of the portfolio, consonant with Title I regulations, required the same content standards and scoring criteria for all students. This was a departure from the dual system used the previous year for English proficient and limited English proficient students. In addition, portfolio components were developed and available for prekindergarten students. The portfolio assessment was not mandated district-wide. Thus, implementation was a school-based decision. As a result, portfolios were implemented in 62% (71 of 118) of the Title I elementary schools. Some schools elected to implement other types of assessment measures for primary students.

Title I evaluation personnel and instructional specialists conducted 26 staff development sessions on portfolio assessment for more than 1,300 teachers. A trainer-of-trainers format allowed school representatives to conduct additional school-level training. Training sessions focused on portfolio assessment procedures described in the teacher's manual, work sample selection guidelines, and use of rubrics to score work samples. In addition to staff development, implementing schools were monitored through school visits to ascertain whether schools were following guidelines and to provide feedback so that teachers could make midcourse adjustments.

The Reading/Language Arts Portfolio Assessment is a comprehensive assessment system designed for students in prekindergarten through second grade that includes four interrelated elements: (a) student work samples, (b) instructional goals and performance criteria, (c) student summary, and (d) scoring rubrics (see Table 1). The assessment system is aligned with content standards, is integrated within classroom instruction, samples a broad range of students' work, and documents literacy development over time.

Student work samples are accumulated over the course of the year by the classroom teacher according to a set of guidelines. First, the student work samples are to align with instructional goals and performance criteria based on the Texas content standards. Second, teachers are encouraged to include assessments or work sam-

TABLE 1
Four Elements of the Reading/Language Arts Assessment

Student work samples	Align with state content standards
	Provide varied, authentic, contextualized learning tasks
	A minimum of 12 work samples is required
	Portfolio Assessment Supplement (Dallas Public Schools, 1996) provides exemplars
Instructional goals/performance criteria	Identify key student outcomes based on state content standards
	Define the specific knowledge/skills students will know and use
	Guide instruction and the selection of work samples aligned with content standards
Student summary	Lists the instructional goals and performance criteria for a particular grade level
	Provides an ongoing record of each student's growth and progress for fall, winter, and spring rating periods
Scoring rubics	Designate a scale ranging from lowest to highest as follows:
	1. Emerging—minimal level of student performance
	2. Developing—evolving, inconsistent level of performance
	3. Proficient—standard all students expected to achieve
	4. Distinguished—outstanding or superior performance
	Establish criteria for how well students should do their work

ples that provide children with meaningful, contextualized learning events and to include items from a variety of media (e.g., writings, reading logs, anecdotal records, performance assessments, etc.). Third, a minimum of 12 student work samples is required. The teacher has flexibility in work sample selection and involving the students in the process; however, the portfolio contents must document the student's level of performance completely. To assist the teacher in the selection of tasks, a *Portfolio Assessment Supplement* (Dallas Public Schools, 1996) is provided with examples of student work samples that align with the content standards.

The instructional goals, based on content standards, identify the key student outcomes that are assessed. Goals are outlined for each grade level (prekindergarten to second grade). Prekindergarten (PK) and kindergarten (K) portfolios are rated on four goals, and first- and second-grade portfolios are rated on six. Each grade-level goal has between three to nine student performance criteria. The performance criteria describe the knowledge and skills that students should know and use, and as a result, serve to inform instructional practice and guide teachers in selecting assessments or work samples that align with the content standard. For example, a checklist identifying the capital and lower case letters a student had mastered could serve as a work sample to meet the kindergarten goal "the student will use print conventions" and the performance criterion "recognize the letters of the alphabet."

The student summary provides an ongoing record of each student's growth and progress toward achieving proficiency at the grade-level standard for three assessment periods (fall, winter, and spring). The summary also provides a profile of a student's accomplishments related to the goals and performance criteria. Relying on scoring rubrics, the teacher makes two types of ratings on the student summary during each period. First, check marks, corresponding to the quality of each student work sample, are recorded beside the pertinent performance criterion. Second, instructional goal ratings are judged at the end of each rating period based on the collection of work samples that align with the criteria for each goal.

The scoring rubrics align with the instructional goals and performance criteria and establish standards for how well students should do their work. The rubrics designate a 4-point scale ranging from 1 (*emerging*) to 4 (*distinguished*). The third point on the scale (*proficient*) represents the standard expected for all students. Scoring rubrics guide teachers' judgments regarding the quality of each piece of work. In addition to scoring individual work samples, teachers make an overall judgment of how well a collection of work samples meets the multidimensional standards that define each instructional goal. The kindergarten scoring rubrics and student summary (including goals and performance criteria) are included in Appendix A and Appendix B, respectively.

METHODOLOGY

First Raters

Classroom teachers scored their own students' portfolios. Instructional goal ratings were recorded on a student summary that was included in each child's portfolio. The teacher ratings used in this analysis were those made during the spring rating period for the four prekindergarten and kindergarten goals and the six Grades 1 and 2 goals.

Second Raters

In May, 42 teachers participated as second raters to re-score a sample of portfolios. Paid participants included prekindergarten ($n = 7$), kindergarten ($n = 12$), Grade 1 ($n = 12$), and Grade 2 ($n = 11$) teachers. Participants were randomly selected from a list of teachers who had attended district portfolio training sessions and had implemented portfolios during the current year. Prior to actual scoring, second raters received additional training that allowed them to review a sample portfolio, discuss the language of the rubrics and the characteristics of the work samples to achieve consensus, and record information for a sample portfolio.

Procedures

A two-stage portfolio sampling process was used. First, a random sample of elementary schools was selected (10 schools in total were sampled). Next, a random sample of approximately 10% of the prekindergarten to second grade portfolios was selected for re-scoring. The student summaries were removed from each portfolio and assigned a tracking number. A blank student summary was placed in each portfolio so that second raters could blindly rate the student's performance.

Over a 2-day period (a total of approximately 252 person hrs), 150 portfolios were judged on the instructional goals. Each second rater independently reviewed a subsample of portfolios for their assigned grade level (e.g., Grade 1 teachers reviewed first-grade portfolios). Raters judged 62 prekindergarten and kindergarten portfolios and 88 Grades 1 and 2 portfolios. Each rater reviewed three or four portfolios, but they never scored portfolios from their own school.

During scoring sessions, second raters recorded information for each reviewed portfolio in two ways. First, instructional goal ratings were made on a student summary form for the spring rating period. When evidence was inadequate to judge the quality of the student's work, a goal received a "no evidence" rating. Second, raters reported information on a response sheet to identify a student's grade level, the number of work samples in the portfolio, whether classroom teachers specified the goal and performance criteria addressed by the sample, the degree to which teacher comments on the work samples provided explanation or clarification about the student's work, and the instructional goals that were documented effectively enough to rate the student's performance.

RESULTS

Reliability of Portfolio Scores

Several measures of classroom teacher and second rater agreement were utilized. These included mean score differences, percentages of agreement, and interrater correlations.

Mean scores for instructional goals.   The information in Table 2 shows the prekindergarten and kindergarten and Grades 1 and 2 mean scores assigned by the teacher and second rater for the portfolio instructional goals across all portfolios that were rated. Scores were analyzed using paired sample t tests. Effect sizes (Cohen's d) were calculated to note practical significance (Cohen, 1988).

Classroom teachers, overall, assigned higher average scores than the second raters for both prekindergarten and kindergarten and Grades 1 and 2 instructional

TABLE 2
Mean Scores for Instructional Goals by Rater Type and Grade Level

Instructional Goal	n^a	Teacher		Second Rater		t Value	Effect Size
		M	SD	M	SD		
Prekindergarten and kindergarten							
Vocabulary/comprehension	36	2.86	.80	2.39	.65	−3.11***	.64
Print conventions	45	2.91	.70	2.38	.83	−4.12***	.69
Write about experiences	36	2.81	.92	2.47	.70	−1.87	.42
Ownership of literacy	26	2.88	.86	2.50	.71	−2.18*	.48
Grades 1 and 2							
Listening and speaking	38	2.58	.89	2.39	.72	−1.36	.23
Vocabulary strategies	54	2.83	.93	2.39	.79	−4.12***	.51
Comprehension strategies	42	2.79	.93	2.43	.74	−2.72**	.43
Study strategies	42	2.98	.95	2.48	.77	−4.37***	.58
Ownership of literacy	41	2.63	.99	2.32	.82	−2.31*	.34
Generate compositions	60	2.57	.87	2.22	.76	−3.31***	.43

Note.   There were 62 prekindergarten and kindergarten portfolios reviewed and 88 Grades 1 and 2 portfolios reviewed. Means and standard deviations were based on a 4-point scale ranging from 1 (*lowest*) to 4 (*highest*).

^{a}Indicates the number of cases in which scores were given by both raters.

*$p < .05$ **$p < .01$ ***$p < .001$.

goals. The higher ratings were statistically significant for all goals except "write about experiences" for prekindergarten and kindergarten and "listening and speaking" for Grades 1 and 2. Effect sizes confirmed the magnitude of the differences between ratings. The discrepancies between raters' average scores may reflect teacher bias; that is, classroom teachers are more lenient in assigning scores. Averaged across goals, the scores assigned by the students' own teachers were approximately 0.40 points (on the 4-point scale) higher than the second raters.

Consistency of instructional goal scores.   Mean instructional goal scores revealed the average scores across all portfolios; however, it was of interest to determine the proportion of cases in which both raters agreed on a score. The percentage of agreement between teachers' and second raters' spring goal ratings are displayed in Table 3. The difference between scores was calculated by subtracting the second rater's rating for a particular instructional goal from the teacher's rating for the same goal. Results in Table 3 indicate the percentage of cases with 0- to 3-point differences and the percentage of instructional goals for which the evidence in the portfolio was inadequate for the second rater to assign a score.

The percentage of agreement between teachers and second raters varied by instructional goal and grade level. A relatively high percentage (50% to 59%) of

TABLE 3
Percentage of Agreement Between Teachers' and Second Raters' Spring Instructional
Goal Ratings by Grade Level

	Ratings Given by Both Raters						
	Rated	% of Difference in Ratings[a]				No Evidence[b]	
Instructional Goal	n[c]	0	1	2	3	n[d]	%
Prekindergarten and kindergarten							
Vocabulary/comprehension	36	47	36	17	0	26	42
Print conventions	45	38	49	13	0	17	27
Write about experiences	36	33	53	11	3	26	42
Ownership of literacy	26	50	42	4	4	36	58
Grades 1 and 2							
Listening and speaking	38	53	40	8	0	50	57
Vocabulary strategies	54	52	37	11	0	34	39
Comprehension strategies	42	52	36	12	0	46	52
Study strategies	42	50	41	10	0	46	52
Ownership of literacy	41	59	27	15	0	47	53
Generate compositions	60	55	37	7	2	28	32

Note. There were 62 prekindergarten and kindergarten portfolios reviewed and 88 Grades 1 and 2 portfolios reviewed. Percentages may not add up to 100 due to rounding.
[a] Percentage of cases with the indicated differences in scores for the two independent ratings. Includes cases in which ratings were given by both raters. [b] Evidence in the portfolios was inadequate for the second rater to assign a score. [c] Number of valid cases in which ratings were given by both raters. [d] Number of cases in which the second rater indicated that there was insufficient evidence to assign a rating.

Grade 1 and 2 portfolios received exactly the same score by the two raters when the evidence was adequate. Identical ratings occurred less frequently (33% to 50%) and less consistently for prekindergarten and kindergarten instructional goals. When raters were not in perfect agreement, a one-point difference was most likely to occur. Across all goals and grade levels, the percentage of one-point differences between raters ranged from 27% to 53%. Two-point differences between raters seldom occurred and three-point discrepancies were rare.

Ratings could not be made about half of the time for many instructional goals due to inadequate evidence. This probably occurred because some portfolio goals were easier to document than others. For example, in prekindergarten and kindergarten, "print conventions" (27%) had the lowest percentage of portfolios with no evidence. A lower percentage of the Grades 1 and 2 portfolios had inadequate evidence for "generate compositions" (32%). Overall, because some degree of agreement would be expected by chance alone and because a large percentage of the instructional goals had inadequate evidence to assign ratings, the reliability of the

portfolio goal ratings was low. If computations had included all cases (with and without adequate evidence), the percentages would be considerably lower.

Interrater correlations.   Another measure of rater reliability is the correlation between raters' scores. The interrater correlation coefficients between the classroom teachers' and second raters' instructional goal ratings are presented in Table 4.

The correlation coefficients for prekindergarten and kindergarten ranged from .16 for "write about experiences" to .40 for "ownership of literacy." There was a low mean correlation of .37 for all goals. The correlation coefficients were somewhat higher for Grades 1 and 2, ranging from .41 for "listening and speaking/comprehension strategies" to .57 for "study strategies." The average correlation was .48. Collectively, the correlation coefficients were low. A mean prekindergarten and kindergarten correlation of .37 indicated that only about 14% of the variance in second raters' scores was explained. Likewise, the first- and second-grade mean correlation of .48 indicated that 23% of the variance in second raters' scores could be predicted from knowing the teachers' scores. Large percentages of unexplained variance were due to error.

TABLE 4
Interrater Correlations Between Classroom Teachers' and Second
Raters' Instructional Goal Ratings by Grade Level

Instructional Goal	n[a]	r
Prekindergarten and kindergarten		
Vocabulary/comprehension	36	.25
Print conventions	45	.35
Write about experiences	36	.16
Ownership of literacy	26	.40
M	36	.37
Grades 1 and 2		
Listening and speaking	38	.41
Vocabulary strategies	54	.51
Comprehension strategies	42	.41
Study strategies	42	.57
Ownership of literacy	41	.47
Generate compositions	60	.48
M	46	.48

Note. There were 62 prekindergarten and kindergarten portfolios reviewed and 88 Grades 1 and 2 portfolios reviewed.
[a] Indicates the number of cases in which scores were given by both raters.

TABLE 5
Second Raters' Responses to "Which Instructional Goals Were Documented Effectively
Enough to Rate the Student's Performance?"

Instructional Goal	Adequately		Inadequately	
	n	%	*n*	%
Prekindergarten and kindergarten				
Vocabulary/comprehension	36	58	26	42
Print conventions	45	73	17	27
Write about experiences	36	58	26	42
Ownership of literacy	26	42	36	58
Grades 1 and 2				
Listening and speaking	38	43	50	57
Vocabulary strategies	54	61	34	39
Comprehension strategies	42	48	46	56
Study strategies	42	48	46	52
Ownership of literacy	41	47	47	53
Generate compositions	60	68	28	32

Note. There were 62 prekindergarten and kindergarten portfolios reviewed and 88 Grades 1 and 2 portfolios reviewed.

Validity of Inferences Made From the Portfolio Scores

Validity based on second raters' analysis. During the portfolio rating sessions, the second teacher raters determined whether the content of the portfolio was sufficient to judge a student's performance. The number of work samples included in the reviewed portfolios showed that some portfolios in prekindergarten and kindergarten (14%) and Grades 1 and 2 (7%) did not have the stipulated minimum of 12 work samples. Portfolios for older students were likely to have more work samples than those for younger students. It was impossible to document students' content knowledge without an adequate number of work samples.

Inadequate documentation for the individual instructional goals represented another content-related problem for the portfolios. As shown in Table 5, second raters judged a large percentage of portfolios as having inadequate documentation to enable them to rate the various goals. Some goals were documented more completely than others. "Ownership of literacy" was adequately documented in only 42% of prekindergarten and kindergarten and 47% of Grades 1 and 2 portfolios. On the other hand, "print conventions" (73%) and "generate compositions" (68%) had adequate work samples more often. Work samples for these goals were probably generated during daily classroom activities.

Another problem that made it difficult for second raters to judge students' performance was the lack of alignment between the work samples in the portfolio and

the instructional goals. On about one-third of the work samples, teachers neglected to record data to designate the instructional goal that was assessed. It was difficult to determine the purpose of the work sample when the instructional goal data were missing.

According to the portfolio implementation guidelines, teachers were encouraged to add specific comments, whenever appropriate, that would evaluate the student's performance. Teacher notations were helpful to explain the instructional activity generating the work sample and to clarify the student's performance. Second raters usually found few, or no, teacher comments on the work samples in the portfolios. Of the portfolios they reviewed, 29% of the prekindergarten and kindergarten and 43% of the Grades 1 and 2 portfolios had no teacher comments. Overall, for many portfolios, there was insufficient sampling of the content domain because the number of samples was inadequate, the work samples provided inadequate information, the purpose for the work samples was unknown, the work samples did not exemplify the goal's content knowledge, or there were no teacher notations to explain activities and to clarify the student's performance.

Second raters' perceptions of the validity of portfolio content. During the portfolio rating sessions, raters wrote notes citing work sample problems that made rating students' performance difficult, and they identified the types of work samples that provided the most valid evidence. Second raters noted that many teachers did not adhere to work sample selection guidelines. For example, work samples often lacked dates and were not aligned with the goals, some goals were not well documented, and there were few teacher comments or explanations. The quality of some work samples was questionable. For instance, raters believed that photocopied worksheets provided little information about students, and it was sometimes unclear whether work samples were teacher modeled or the students' own work.

Second raters also identified the types of work samples that provided valid evidence about students. Prekindergarten and kindergarten teachers, for example, cited the importance of using anecdotal records, writings with invented spelling, and one-to-one assessments to document the progress of young children. Grades 1 and 2 teachers emphasized the importance of authentic reading tasks, informal assessments, and varied types and stages of student-generated writing to reveal the literacy accomplishments of older students.

Criterion-related validity of instructional goals. Due to low interrater reliability and content sampling limitations, the review of concurrent validity was for exploratory purposes only. Convergent and divergent evidence was examined by determining the relation between students' spring portfolio goal ratings (1 to 4) and relevant scores on the Iowa Tests of Basic Skills (ITBS) reading, language, and mathematics subtests.

For kindergarten students, convergent evidence was examined with the ITBS word analysis and vocabulary subtests, while divergent evidence was reviewed for the ITBS mathematics subtest. The word analysis subtest assessed letter recognition and letter-sound relations, whereas the vocabulary subtest measured listening vocabulary. The mathematics subtest consisted of questions presented orally to measure beginning mathematics concepts, problem solving, and math operations.

Convergent evidence for Grades 1 and 2 was studied with the ITBS reading comprehension, reading total, and language total subtests, whereas divergent evidence was examined with the ITBS math computation and mathematics total subtests. Reading comprehension assessed factual, inferential, and evaluative meaning. Reading total included the same comprehension strategies plus vocabulary items. Word attack items were included for first grade only. The language subtest measured students' abilities to understand linguistic relations and usage. Mathematics total included concepts, problem solving, data interpretation, and computation, whereas the computation subtest included only mathematical computation. Only a few schools elected to administer the language subtest; thus, the sample size was small, and the correlations are representative of that limited subsample of students.

The correlations between the kindergartners' goal ratings and the ITBS subtests showed low positive relations ranging from .32 to .52 (see Table 6). As seen in Table 7, the correlations for Grades 1 (.42 to .54) and 2 (.38 to .59) were somewhat higher than kindergarten correlations. In all cases, no definitive convergent or divergent relation patterns were present. The weak association between portfolio scores and the standardized assessments does not necessarily indicate that the portfolio scores are invalid. It is possible that portfolio assessment, stressing both product and process, measures aspects of reading and language arts that standardized assessments do not measure.

Some of the low correlations were not surprising. For example, the skill-based language subtest used with first and second grade was probably not a suitable comparison with the portfolio goals' emphasis on language and writing processes.

TABLE 6
Correlations Between Kindergarten Portfolio Instructional Goal Scores and ITBS Raw Scores for Word Analysis, Vocabulary, and Mathematics

Instructional Goal	ITBS Reading/Language Arts			ITBS Math	
	n	Word Analysis	Vocabulary	n	Total
Vocabulary/comprehension	3,916	.51	.37	4,138	.49
Print conventions	3,913	.52	.33	4,135	.49
Write about experiences	3,874	.48	.33	4,094	.47
Ownership of literacy	3,811	.47	.32	4,031	.46

Note.   ITBS = Iowa Tests of Basic Skills.

TABLE 7
Correlations Between Grade 1 and 2 Portfolio Instructional Goal Ratings and ITBS Raw Scores for Reading, Language, and Mathematics

Instructional Goal	n[a]	Reading		Language	Mathematics	
		Comprehension	Total	Total[b]	Computation	Total
Grade 1						
Listening and speaking	2,097	.50	.49	.42	.44	.51
Vocabulary strategies	2,088	.54	.53	.44	.45	.51
Comprehension strategies	2,079	.54	.54	.43	.46	.53
Study strategies	2,023	.52	.51	.44	.46	.52
Ownership of literacy	2,105	.52	.51	.44	.44	.51
Generate compositions	2,105	.50	.48	.45	.43	.50
Grade 2						
Listening and speaking	2,469	.56	.55	.52	.38	.49
Vocabulary strategies	2,450	.59	.59	.53	.40	.51
Comprehension strategies	2,462	.59	.59	.52	.41	.52
Study strategies	2,421	.58	.58	.54	.43	.53
Ownership of literacy	2,484	.56	.56	.50	.42	.49
Generate compositions	2,484	.56	.55	.45	.42	.49

Note.   ITBS = Iowa Tests of Basic Skills.
[a]The mean sample size for ITBS reading and mathematics tests. [b]A limited number of schools elected to administer the ITBS language subtest. The mean sample size for Grade 1 language was 326 and Grade 2 language was 592.

Similarly, comparable correlations between kindergarten goal ratings and mathematics and reading/language arts subtest scores were not unexpected because the mathematics test consisted of questions presented orally. For Grades 1 and 2, there was a stronger association between the portfolio goals and the mathematics total subtest (which encompassed concepts, problem solving, interpretation, and computation) than between the goals and the mathematics computation subtest (which focused exclusively on mathematical computing). For all grades, there appeared to be confounding effects when mathematics tests required students to use reading and language skills to interpret data and to solve written problems.

SUMMARY

Reliability was low for portfolio scores at all grade levels because interrater reliability was low, and the varying tasks and evidence within the portfolios influenced the reliability of scores. Second raters often found it difficult to rate student performance based on the varying tasks and evidence in the portfolios. Teachers' portfolio ratings were systematically higher than second raters' across most instructional

goals. Whereas the discrepancies between raters' average scores may reflect actual teacher bias, the differences may also be due to the insufficient evidence in the portfolios for generating accurate independent ratings. When evidence in the portfolio was adequate, there was a notable percentage of exact, or close, agreement between teachers and second raters. However, many instructional goals had inadequate evidence to assign ratings. Interrater correlations for the instructional goals were low, ranging from .37 for prekindergarten and kindergarten to .48 for Grades 1 and 2, and these correlations were inflated because cases were dropped when only one score was available.

When second raters judged the adequacy, or validity, of the content of the portfolios for making ratings, inferences about students' performance were undermined for several reasons. First, a large percentage of portfolios were judged to have inadequate documentation to rate the various goals. Second, some goals were documented more completely than others. Third, the absence of data to link the work sample to a particular instructional goal often made the purpose of student tasks unclear. Last, teacher comments clarifying students' performance were scant.

No definitive concurrent relations were established. As for convergent evidence, there were low positive correlations between the portfolio instructional goal ratings and closely related ITBS subtest scores. Similarly, low correlations were found between students' portfolio mastery status and ITBS scores. Nonetheless, it is possible that portfolio assessment, stressing both product and process, measures aspects of reading that standardized assessments do not measure. Divergent relations with mathematics were not firmly established. The differences between the mathematics subtest score associations pointed to confounding effects of reading and language development when mathematics tests required students to read and solve written problems and to interpret data. Mathematics computation showed a lower correlation because only mathematical computing was involved. This confounding element partly explained the lack of divergent validity.

CONCLUSIONS AND IMPLICATIONS

It seems that more schools are using portfolios in elementary schools, yet little evidence exists related to the technical quality of this form of assessment in the primary grades. Portfolios for younger children generally do not involve high-stakes accountability decisions. Nevertheless, if portfolios provide the basis for decision making about students, then attention to the quality of scores is warranted. The Reading/Language Arts Portfolio Assessment aimed to achieve the dual purposes of measuring students' progress toward state content and performance standards and influencing instructional decision making. To that end, the portfolio relied on a structural framework but allowed teacher flexibility in the selection of work samples. Thus, the portfolio was unstandardized. Findings from the study of the portfo-

lio assessment have implications for others who are designing literacy assessments for young children. Three issues emerged from the results: expectations regarding technical quality, implications for teachers and training, and expectations for cost and effort.

Expectations for Technical Quality

After 3 years of development, the portfolio assessment did not provide high quality information about student achievement for either instructional or informational purposes. The unreliability of the scores was likely related to (a) lack of standardization of tasks, (b) problems with the scoring rubrics, and (c) inadequate training. Specifically, the portfolios contained a broad range of work samples, some of which were inappropriate to document the goals. The diversity of tasks complicated the scoring process and restricted comparisons among students. Because the tasks were unstandardized, the scoring rubrics aligned with the instructional goals rather than with specific tasks. As a result, the rubrics described student performance expectations in very general terminology that likely contributed to the inconsistent scoring. Improving the scoring rubrics will require greater standardization of the portfolio contents so that there is stronger alignment between tasks and specific evaluative criteria. This suggests the need for a compromise between standardization, which is needed to improve technical quality, and the flexibility that allows portfolios to be integrated within the classroom context. In addition to standardization and high-quality rubrics, reliable scoring is dependent on effective training. Unfortunately, the training provided did not give classroom teachers and second raters sufficient opportunities to understand the portfolio process and to apply the scoring criteria.

Problems with validity reflected an inadequate sampling of the reading/language arts domain. Some goals were well documented. Other instructional goals had insufficient evidence to permit judgments. To improve validity and reliability, it seems that portfolios need to contain a core of *essential* work samples (those that all portfolios must contain) and *optional* work samples (those that the teacher and students are free to select). The core samples would provide a common frame of reference across all portfolios for judging students' performance.

Implications for Teachers and Training

It was believed that portfolios offered a means to build teachers' knowledge about a quality literacy curriculum for young children; however, reality suggested a need for initiatives to reform instruction as well as assessment. In short, a portfolio was only as good as the teacher who constructed it and the learning opportunities that were provided for the students. Although developmentally appropriate tasks were described during training and provided in the portfolio supplement, evidence

showed that many portfolios had inadequate numbers of work samples or were filled with work samples that did not provide worthwhile learning experiences for children (e.g., worksheets, copying, coloring). This suggests that unless extensive staff development is provided, literacy portfolios comprised of random collections of work samples may be unrealistic for large-scale assessment systems that seek to provide scores that accurately describe student proficiency.

Expectations Regarding Cost and Effort

This study indicates that high cost and effort are major obstacles to the development and use of early childhood portfolios on a large-scale basis. Portfolio assessment systems require time for development and implementation as well as dollars for materials and personnel. Assembling the assessment portfolios was a time-consuming task for teachers because they had to attend staff development, interact with students, select work samples, record documentation on work samples, score work samples, and record information on the student summary. Evidence accumulated during the reliability study indicated that many teachers failed to adhere to the assessment guidelines. This showed that teachers either needed additional training to understand the process, or they did not have the time or motivation to fulfill all of the requirements. In either case, the quality of the evidence in the portfolio was compromised.

Altogether, the findings of this study raise cautions to those who believe portfolios are the answer to the perceived ills of standardized testing for young children. When portfolios are used for classroom instruction and low-stakes informational purposes, they require a long, difficult, and expensive development process as well as some degree of standardization to provide credible evidence about students' achievement. School districts must decide whether they are willing to commit the necessary time and money to develop a technically adequate portfolio assessment system. Otherwise, the assessment will have limited value.

Dallas' portfolio assessment did not meet adequate and hoped-for technical standards. However, the portfolios did provide a means to communicate about valued instructional goals, methods, and outcomes. It was encouraging to see the portfolio evolve from a folder containing work samples to a more structured assessment system that improved gradually over the 3 years of implementation. Even so, much remains to be done before a high-quality portfolio assessment is a reality.

REFERENCES

Arter, J. A., & Spandel, V. (1992). Using portfolios of student work in instruction and assessment. *Educational Measurement: Issues and Practice, 11*(1), 36–44.
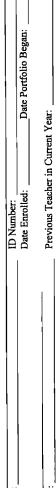
Boykin, W., Yang, H., & Benoit, J. (1993). *Final evaluation of the 1992–93 Chapter 1 instructional program* (PEA93–230–05). Dallas, TX: Dallas Independent School District, Division of Program Evaluation and Accountability Services.

Cohen, J. (1988). *Statistical power analysis* (2d ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Dallas Public Schools. (1995). *Local education agency plan: 1995–2000.* Dallas, TX: Dallas Public Schools, Office of Title I Instructional Program.

Dallas Public Schools. (1996). *Portfolio assessment supplement.* Dallas, TX: Dallas Public Schools, Division of Research, Planning, and Evaluation.

The Evaluation Center. (1995). *An independent evaluation of the Kentucky Instructional Results Information System (KIRIS).* Kalamazoo: Western Michigan University.

Frechtling, J. A. (1991). Performance assessment: Moonstruck or the real thing? *Educational Measurement: Issues and Practice, 10*(4), 23–25.

Gearhart, M., & Herman, J. L. (1995, Winter). *Portfolio assessment: Whose work is it? Evaluation comment.* (Available from the UCLA Center for the Study of Evaluation, 10880 Wilshire Boulevard, Suite 700, Los Angeles, CA 90024–1394)

Gearhart, M., Herman, J. L., Baker, E. L., & Whittaker, A. (1993). *Whose work is it? A question for the validity of large-scale portfolio assessment* (Tech. Rep. No. 363). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing, Center for the Study of Evaluation.

Herman, J. L., Aschbacher, P. R., & Winters, L. (1992). *A practical guide to alternative assessment.* Alexandria, VA: Association for Supervision and Curriculum Development.

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

Herman, J. L., & Winters, L. (1994). Portfolio research: A slim collection. *Educational Leadership, 52*(2), 48–55.

Improving America's Schools Act of 1994, Pub. L. No. 103–382, § 198 Stat. 3518 (1994).

Koretz, D., Stecher, B., Klein, S., & McCaffrey, D. (1994). The Vermont portfolio assessment program: Findings and implications. *Educational Measurement: Issues and Practice, 13*(3), 5–16.

Koretz, D., Stecher, B., Klein, S., McCaffrey, D., & Deibert, E. (1993). *Can portfolios assess student performance and influence instruction?* (Tech. Rep. No. 371). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

LeMahieu, P. G., Gitomer, D. H., & Eresh, J. T. (1995). Portfolios in large-scale assessment: Difficult but not impossible. *Educational Measurement: Issues and Practice, 14*(3), 11–28.

Linn, R. L. (1994). Performance assessment: Policy promises and technical measurement standards. *Educational Researcher, 23*(9), 4–14.

Mehrens, W. A. (1992). Using performance assessment for accountability purposes. *Educational Measurement: Issues and Practice, 11*(1), 3–20.

Meisels, S. J. (1997). Using work sampling in authentic assessment. *Educational Leadership, 54*(4), 60–65.

Reckase, M. D. (1995). Portfolio assessment: A theoretical estimate of score reliability. *Educational Measurement: Issues and Practice, 14*(1), 12–31.

Shapley, K. S., & Pinto, M. F. (1995). *Final report of the development of the 1994–95 Chapter 1: Portfolio assessment* (REIS95–274–2). Dallas, TX: Dallas Public Schools, Division of Research, Planning, and Evaluation.

Sheehan, D., Yang, H., Shapley, K., Johnson, A., & Thapa, M. (1994). *Final evaluation of the 1993–94 Chapter 1: Instructional program* (REIS94–270–2). Dallas, TX: Dallas Independent School District, Division of Research, Planning, and Evaluation.

Tierney, R. J., Carter, M. A., & Desai, L. E. (1991). *Portfolio assessment in the reading-writing classroom.* Norwood, MA: Gordon.
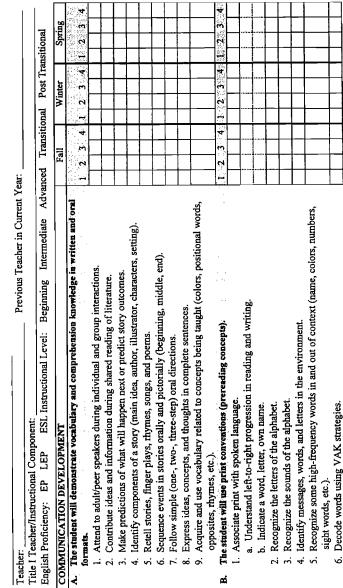
Kindergarten Scoring Rubics[1]

PORTFOLIO SCORING RUBRIC[1]
COMMUNICATION DEVELOPMENT - KINDERGARTEN

Emerging 1	Developing 2	Proficient 3	Distinguished 4
The student will demonstrate vocabulary, concepts and comprehension knowledge in written and oral formats.			
Seldom attends to speakers. Rarely participates during shared reading. Seldom attends to the message conveyed in texts. Not yet able to utilize picture and meaning clues to make predictions. Unaware of story components. Unable to retell stories, etc. Events seldom sequenced in logical order. Not yet able to follow simple oral directions. Expresses ideas using words and phrases. Limited understanding of vocabulary concepts.	1. Limited attention to speakers. 2. Occasionally contributes ideas and information during shared reading. Limited attention to the text's meaning. 3. Relies on teacher prompts to make predictions about events and meanings in stories. 4. Identifies some story components. 5. Retells with teacher prompts. 6. Sequences events with teacher prompts. 7. Follows one-, two-step directions. 8. Beginning to use some complete sentences. 9. Developing understanding of vocabulary being taught.	1. Attends to speakers' message. 2. Often contributes ideas and information during shared reading. Listens and demonstrates understanding. 3. Generally uses title, picture, and meaning clues to make and justify predictions. 4. Identifies components of a story. 5. Retells stories, etc. accurately. 6. Logically orders a basic sequence of events. 7. Follows three-step directions. 8. Clearly describes or discusses ideas in complete sentences. 9. Using acquired vocabulary appropriately.	1. Actively attends to speakers. 2. Insightful contributor to shared reading. Listens and analyzes the text's message. 3. Makes predictions using title, picture, and meaning clues. Listens to confirm or deny predictions. 4. Identifies story components and relates to other contexts. 5. Retells stories, etc. in detail. 6. Identifies and sequences the main events in stories, etc. 7. Follows and interprets complex directions. 8. Communicates in complex sentences. 9. Uses a well developed vocabulary.
The student will use print conventions (prereading concepts).			
Unaware of the functions of print conventions. Little or no knowledge of the letters of the alphabet. Little or no knowledge of the sounds of the alphabet. Rarely is aware of the presence and meaning of environmental print. Unable to recognize high-frequency words. Not yet able to use phonetic strategies to decode words.	1. Developing knowledge of print conventions. 2. Identifies some letters of the alphabet. 3. Identifies some sounds of the alphabet. 4. Occasionally recognizes and gains meaning from print in the environment. 5. Sometimes recognizes high-frequency words. 6. Uses phonetic strategies with teacher support.	1. Understands the functions of print conventions in reading and writing. 2. Identifies the letters of the alphabet. 3. Identifies the sounds of the alphabet. 4. Frequently identifies and understands letters, words, and messages in the environment. 5. Recognizes targeted high-frequency words in and out of context. 6. Independently uses phonetic strategies to decode words.	1. Uses knowledge of print and the predictability of language to gain meaning from texts. 2.-3. Uses knowledge of letters/sounds to identify words while reading/writing. 4. Consistently recognizes letters, words, messages in the environment. 5. Consistently identifies high-frequency words. 6. Uses phonetic strategies while reading/writing independently.
The student will write about experiences.			
1. Limited ability to share ideas orally and to narrate events from personal experience. 5. Scribble stage. Scribbles and pretends to write. No recognizable letters. Writing is represented as lines, scribbles, and scrawls.	1-3. Developing ability to share ideas and narrate personal experiences. 3-5. Isolated letter stage. • Uses letters or letter-like signs to represent writing. • There is an understanding of the purpose for writing. Some drawings help the writer hold the meaning and read back over time.	1-3. Confidently communicates ideas and personal experiences. 3-5. Transitional stage. • Some conventionally spelled words are used correctly. • Writing often returns to isolated letters, symbols, and numerals. • Writes own name, copies words including all letters, and can invent the spelling of words using phonetic clues.	1-3. Excels in contributing ideas during discussions and writes experiences independently. 3-5. Stylized sentence stage. • Creates sentences around known words, repetitive phrases, and sentence beginnings. • Uses words from the environment to complete sentences. • The writer and others can read the message encoded in print.
The student will exhibit ownership of literacy.			
Not yet interested in books or stories. Few or no literary preferences. Rarely participates in shared reading/writing. Little interest in selecting books or taking books home to share. Rarely evaluates own work, learning, or progress.	1. Limited interest and attention to print, stories, books. 2. Interested in picture books and some stories. 3. Sometimes participates in shared reading/writing. 4. Limited interest in books. 5. Seldom takes initiative to evaluate own work, learning, or progress.	1. Usually enjoys listening to stories and exhibits interest in print and books. 2. Often indicates interest in favorite books and stories. 3. Active and successful participant in shared reading/writing. 4. Often selects favorite books and wants to take books home. 5. Frequently assesses own work, learning, and progress.	1. Almost always enjoys listening to stories and shows real interest in books and print. 2. Returns many times to favorite books. Wants to retell and share favorites. 3. An insightful contributor to shared reading/writing. 4. Eager to select/share favorite books. 5. Consistently self-evaluates.

Some ideas related to scoring rubics for writing were adapted from Trail's (1993) *Stages of Writing.*

Portfolio Assessment
Student Summary - Kindergarten

Dallas Public Schools

Student: _____   ID Number: _____
School: _____   Date Enrolled: _____   Date Portfolio Began: _____

Teacher: _____
Previous Teacher in Current Year: _____
Title I Teacher/Instructional Component: _____
English Proficiency: EP LEP ESL Instructional Level: Beginning Intermediate Advanced Transitional Post Transitional

COMMUNICATION DEVELOPMENT	Fall				Winter				Spring			
	1	2	3	4	1	2	3	4	1	2	3	4

A. The student will demonstrate vocabulary and comprehension knowledge in written and oral formats.
1. Attend to adult/peer speakers during individual and group interactions.
2. Contribute ideas and information during shared reading of literature.
3. Make predictions of what will happen next or predict story outcomes.
4. Identify components of a story (main idea, author, illustrator, characters, setting).
5. Retell stories, finger plays, rhymes, songs, and poems.
6. Sequence events in stories orally and pictorially (beginning, middle, end).
7. Follow simple (one-, two-, three-step) oral directions.
8. Express ideas, concepts, and thoughts in complete sentences.
9. Acquire and use vocabulary related to concepts being taught (colors, positional words, opposites, rhymes, etc.).

B. The student will use print conventions (prereading concepts).
1. Associate print with spoken language.
   a. Understand left-to-right progression in reading and writing.
   b. Indicate a word, letter, own name.
2. Recognize the letters of the alphabet.
3. Recognize the sounds of the alphabet.
4. Identify messages, words, and letters in the environment.
5. Recognize some high-frequency words in and out of context (name, colors, numbers, sight words, etc.).
6. Decode words using VAK strategies.

Portfolio Assessment
Student Summary - Kindergarten (Continued)

Dallas Public Schools

Student: _____

COMMUNICATION DEVELOPMENT	Fall	Winter	Spring
1. Participate by expressing ideas/feelings during shared writing.			
2. Dictate ideas and experiences.			
3. Use pictures and other media to describe experiences.			
4. Use letter symbols to form words.			
5. Use letters/words from the environment as a reference when writing.			
6. Recognize and spell some high-frequency words accurately.			
1. Show interest in illustrations, environmental print, stories, and books.			
2. Exhibit preferences for favorite topics, rhymes, stories, books, etc.			
3. Participate confidently in shared reading and writing.			
4. Select books for individual needs.			
5. Self-evaluate own work, learning, and progress.			

Fall	Winter	Spring	⇨Rating periods
			⇨ Circle the appropriate developmental level at the end of each rating period.
✓			⇨ Place a ✓ in the appropriate rating box next to the performance criterion.
✓			⇨ Place a ✓ in the appropriate rating box next to the performance subcriterion.

→ 1 = Emerging   2 = Developing   3 = Proficient   4 = Distinguished

Write on the work sample

• the date
• the instructional goal and performance criteria addressed by the work sample (e.g., Goal A1; Goal C1,3)
• specific teacher comments that evaluate the student's performance

---

[2]The excerpts shown in this appendix are from the portfolio manual developed in the Dallas, Texas Public Schools, 1996.