



# Teaching Portfolios Used for High-Stakes Decisions: You Have Technical Issues!

*Ronald A. Berk*

**DISCLAIMER:**

This chapter, unlike the presentation upon which it is based, does not contain any music. The presentation included the theme from *Mission: Impossible*, “She’s a Lady” (Tom Jones), “Do You Hear the People Sing” (*Les Misérables*), and “YMCA” (The Village People). You’re welcome to hum these selections or some of your own as you plod through this content. In fact, musical accompaniment is highly recommended when reading about portfolios.

You’ve reached your soon-to-be favorite chapter in this book. Based on the volume of feedback I have received on the Q & A format used for the past few years in the NES conference proceedings, I decided to continue that structure here. If you read a question and think, “I don’t care about this,” just skip it until you come to a question you want answered. That should help you whiz through the chapter.

As always, if any content in this chapter seems illogical, inaccurate, or stupid, only one person should be held totally accountable. That person, of course, is Wolf Blitzer.

**Q1: What is a teaching portfolio?**

**A1:** What a great question! Unfortunately, I will not answer that until I've defined the popular concept of portfolio. The teaching portfolio used for high-stakes decisions is a rare species within the broad genus of portfolio, under the family of performance-assessment tools, which is subsumed under the phylum of measurement stuff (according to the Linnaean system of classification).

**Q2: Okay, then, what's a portfolio?**

**A2:** *Portfolio* is derived from two Latin words, "port," meaning "carry," and "folio," meaning "wheelbarrow of material to a powerful person with the hope of scoring higher than I did on the multiple-choice test." Seriously, the most widely quoted definition is by Arter and Spandel (1992) relating to student portfolios: "a purposeful collection of student work that tells the story of the student's effort, progress or achievement in [a] given area[s]" (p. 36). It potentially captures a rich array of what one knows and can do and portrays the processes by which the work is done (Gilman & McDermott, 1994).

**Q3: What distinguishes the portfolio from other types of assessment tools?**

**A3:** After more than a decade of applications of portfolio assessment to measure student and teacher performance, several characteristics have emerged that distinguish it from other forms of "authentic" or "performance" assessment. First, the intent and spirit of the portfolio is to measure outcomes not indicated by paper-and-pencil tests; for example, reflective ability, self-assessment, and personal and professional growth. Second, portfolio assessments contain only "natural" ingredients or elements, that is, actual work over time, settings, and content areas. The portfolio is not artificially created for assessment purposes like a multiple-choice test. Third, the assessment is an on-going process, not a single snapshot in time. The elements are collected over a designated period of time to measure progress toward the outcomes.

**Q4: Is this a new assessment tool?**

**A4:** No, although educators in the 1990s conveyed the impression that a unique and innovative assessment device had hit the educational measurement marketplace. *Au contraire!* (French expression meaning "my shorts are on backwards") The portfolio is an extension of the "work sample" that has been used in business and industry as a method of performance appraisal for more than 50 years (Adkins, 1947), and



the research on its effectiveness and psychometric quality has been conducted in the field of industrial/organizational psychology (Asher & Sciarrino, 1974; Siegel, 1986). The term “portfolio” was borrowed from the visual arts, architecture, and modeling, and applied to student coursework and faculty teaching evidence (Knapper, 1995).

**Q5: How has the portfolio been used in education?**

**A5:** It has served both as a teaching device and as an assessment tool. Probably more than 90 percent of the applications and the research have focused on *students* K–12 (Arter, 1999; Gearhart & Herman, 1995; Gilman & McDermott, 1994; Hall & Hewitt-Gervais, 2000; Herman, Gearhart, & Baker, 1993; Koretz, 1998; Koretz, Stecher, Klein, McCaffrey, 1994; Le Mahieu, Gitomer, & Eresh, 1995; Shapley & Bush, 1999). There is also a growing number of student applications at the college level. For example, portfolio assessment has been implemented at Alverno College, Evergreen State College, King’s College, Miami University, Purdue University, and Southern Illinois University, to name a few. Portfolios are also being used for professional education in medicine (Al Shehri, 1995; Bashook & Parboosingh, 1998; Davis et al., 2001; Friedman Ben-David et al., 2001) and nursing (Brown, 1995; Jasper, 1995). However, despite all of the experience with student portfolios to date, sparse empirical evidence exists to support the frequently cited benefits and the reliability and validity of portfolio scores.

The applications and research on *teaching portfolios* used for high-stakes decisions have been even more limited to the following:

- a. *University faculty*, where the portfolio (a.k.a. *dossier*) (Knapper, 1995) is one of several measures of performance used for promotion and tenure decisions (Berk, 1999; Cox, 1995; Murray, 1995; Seldin, Annis, & Zubizarreta, 1995);
- b. *National Board for Professional Teaching Standards* (NBPTS), which requires a portfolio in conjunction with the assessment center for teacher certification decisions (Educational Testing Service, 1999);
- c. *Connecticut State Department of Education* (CSDE), which is the only state-level application where a portfolio is required for a “Provisional Educator Certificate” decision, the second tier in a three-step certification ladder.



**Q6: Now can you describe the *teaching portfolio*?**

**A6:** Sure. The preceding three applications of teaching portfolios place several restrictions on the components and how they are evaluated. The primary reason for this is due to the high-stakes decisions for which they are used.

Consider the following definition of a teaching portfolio:

*A purposeful collection of direct evidence of teaching practices and reflective/analytic commentary on that evidence that is standardized according to selection criteria, contents, and scoring.*

This definition suggests several key characteristics:

- ▶ All evidence is grounded in teaching practices.
- ▶ Selection criteria and contents are standardized.
- ▶ The teacher selects actual evidence and writes all commentaries.
- ▶ Scoring is standardized according to explicit scoring rubrics and behaviorally anchored scales; concrete benchmarks or performance at each score level; trained scorers; multiple, independent scorers per element; and score paths to arrive at a final score.
- ▶ All procedures are communicated to teachers so that they know all of the rules and what is expected of them.

**Q7: What *elements* can be used in a teaching portfolio?**

**A7:** The starting pool of possible elements could include any of the following:

- ▶ Videotape of teaching
- ▶ Videotape of interactions with students
- ▶ Lesson plans/instructional materials/assignments/daily logs
- ▶ Student work samples with teacher comments
- ▶ Reflective analysis/philosophy
- ▶ Work with colleagues and professional organizations

**Q8: How do you select the elements to include in a specific portfolio?**

**A8:** There are four possible selection criteria: (a) representative, (b) typical, (c) best, or (d) most diverse. Although picking elements that are representative of teaching practices would seem to provide an appropriate *sample*, the NBPTS requests evidence of “best” practices and



the CSDE requires teachers to submit “typical” and “representative” evidence.

**Q9: Are all elements weighted equally in the assessment?**

**A9:** No, not all elements in a portfolio have the same influence on the final decision. Some are considered more important than others. A judgmental panel should be assembled and convened to determine the appropriate weight for each element. Assigning weights is based on reason and values. Although the weights are expressed quantitatively, they are not determined mathematically. The process is entirely judgmental. Factors that the panel should consider are:

- a. authenticity of the evidence of teaching practice;
- b. criticality (or importance) of the skills and abilities being measured; and
- c. breadth and depth of range of skills and abilities sampled.

**Q10: How in the world do you score this puppy?**

**A10:** I have no idea, but I’m going to suggest a strategy. There are two key steps prior to the actual scoring: (1) choose a standardized set of performance anchors and a quantitative scale and (2) explicitly define those anchors as a behaviorally anchored performance scale.

First, select a set of performance anchors (descriptive words) along with a number scale that can be used to evaluate ALL elements in the portfolio. The criteria for scoring must be standardized across all elements. Two sample scales are given below:

3 = Outstanding	2 = Acceptable	1 = Minimally Acceptable	0 = Unacceptable/Putrid
-----------------	----------------	--------------------------	-------------------------

OR

2 = Strong Performance	1 = Acceptable Performance	0 = Needs Improvement
------------------------	----------------------------	-----------------------

Second, write detailed behavioral descriptions for each anchor and portfolio element. This is the portfolio scoring protocol version of what has been referred to as a scoring rubric elsewhere. A rubric is essentially being developed for each element. Suppose you selected the following five elements:

- 1. Videotape I
- 2. Videotape II
- 3. Lesson Plans
- 4. Student Work Samples
- 5. Professional Growth



The task is to describe what “Outstanding” performance looks like for each videotape as well as the other elements. The remaining anchors should be defined accordingly. The reliability of the judgmental scoring of a teaching portfolio hinges on these descriptions. If they are ambiguous, the judges may be confused and more subjective in their scoring, which can decrease interscorer reliability. The above scoring rubric provides a behaviorally anchored performance scale to rate any portfolio containing the required elements.

**Q11: Is there a structure or scoring template that can be used for each candidate and portfolio?**

**A11:** Funny you should ask. A scoring summary template is displayed in Table 1. It includes all elements, performance anchors, judges’ scores, and the weights assigned to the elements. It can produce all of the quantitative information needed to make a pass-fail decision.

**Table 1**  
**SCORING SUMMARY TEMPLATE**

ELEMENT	Outst. 3	Accept. 2	Min. Accept. 1	Unaccept. 0	Aver. Score	Weight	Weighted Score
1.							
2.							
3.							
4.							
5.							

**Q12: Can you demonstrate how to summarize all of the scores?**

**A12:** I thought you would never ask. Table 2 provides the results of the portfolio scoring for Bucko Berk. Let’s begin to decipher the table by moving across the columns from left to right. First, the *elements* in the portfolio are listed. Across the four *anchors* are the scores by two judges for each element. For example, element 2 received ratings of 3 and 2 by two independent judges. Should there be a discrepancy between the judges of 2 points or more, a third judge may be included and then added to the table. The *average score* column simply averages the judges’ ratings. The *weight* column lists the quantitative “relative” weights assigned to the elements by the judgmental panel (see A10). The *weighted score* is the product of the average score and the weight for each element. The scores for all elements are then summed.



This scoring procedure is based on a *compensatory scoring model* where low scores on one or more elements can be compensated by high scores on others. In other words, a candidate’s strengths and weaknesses across portfolio elements are combined. No minimum score is required on any element. The primary argument for choosing this scoring strategy is that the reliability of any single element is usually not sufficiently high enough to justify a high-stakes decision about a teacher. Further, the complexity of teaching suggests that no single element can capture and assess the full range of skills and abilities that teaching requires. A full picture of each candidate’s teaching practices can be obtained by combining as much information together as possible in the complete portfolio. The alternative conjunctive scoring model is less appropriate in that it requires a separate performance standard for every element.

**Bucko Berk**  
Name

**Table 2**  
**SAMPLE SCORING SUMMARY (2 SCORERS)**

ELEMENT	Outst. 3	Accept. 2	Min. Accept. 1	Unaccept. 0	Aver. Score	Weight	Weighted Score
1. Videotape I	3/3				3	10	30
2. Videotape II	3	2			2.5	10	25
3. Lesson Plans	3	2			2.5	8	20
4. Student Work Samples		2	1		1.5	8	12
5. Prof. Growth			1/1		1	5	5

92

Score Range = 0 → 123  
Pass ≥ 82  
Adjusted Pass ≥ 88  
Decision = Pass

On the bottom left under the table, the key summary score information is provided along with the decision. The interpretable score range is calculated based on an average score of 0 for all elements at one extreme and an average of 3 for all elements at the other. The passing score of 82 was determined by a separate standard-setting panel. The adjusted pass, which is optional, accounts for one standard error of 6 points to decrease the probability of false positive errors (those candidates who may pass, but don’t possess the necessary competencies). Finally, the weighted score of 92 is compared with the adjusted passing score of 88 to reach a final decision of “Pass.” It looks like Bucko squeaked through.



**Q13: Should an oral defense of the portfolio be considered before making the final decision?**

**A13:** Absolutely. It may be *required* if the number of candidates does not create an unreasonable burden in terms of time constraints and the availability of oral defense committee members. The defense may be *conditional* based on the candidate's score. If it is slightly above or below the passing score, the defense can provide the opportunity to raise questions about the individual elements, particularly those on which scorers may have disagreed. Clarification by the candidate may be critical to the final decision. The defense may also be *optional*, at the discretion of the candidate.

**Q14: Technically speaking, are there specific types of validity evidence that need to be collected for portfolios?**

**A14:** Yes! They are analogous to most forms of assessment. In portfolio language, the most important evidence relates to the following questions:

1. *Content-Related Validity*

- ▶ Are the *standards/outcomes* being measured by the portfolio representative of effective teaching practices?
- ▶ Does *each element* in the portfolio measure a standard/outcome of effective teaching practice?
- ▶ Are the *elements* in the portfolio a sample of representative, typical, best, or most diverse teaching practices?
- ▶ Do the *scoring criteria* (rubrics) and benchmarks for the elements represent effective teaching practices?

2. *Predictive Validity*

- ▶ Does passing performance on the portfolio predict future effective teaching practices?

3. *Construct Validity*

- ▶ Does portfolio performance correlate with other measures of teaching performance?

4. *Decision Validity*

- ▶ Is the pass-fail decision made about a teacher based on his or her portfolio score accurate?





**Q15: Are there similar criteria for collecting *reliability evidence*?**

**A15:** Of course, but it's probably most appropriate to express the evidence in terms of sources of measurement error. Consider the following questions related to five categories of measurement error:

1. *Assessment Time* – when the portfolio elements are collected.
  - a. Time Sampling
    - ▶ During what period of time were the elements collected?
    - ▶ Are different points of time represented to show growth or just one snapshot in time?
  - b. Performance Reliability
    - ▶ Can performance on the elements be generalized to other points in time?
2. *School setting*—where the portfolio elements are collected.
  - ▶ Is the evidence of teaching site-specific or could it be generalized to multiple school settings?
3. *Element Sampling*—what the portfolio measures.
  - ▶ Is performance on one sample of elements generalizable to a different sample?
4. *Scorer Sampling*—who scored the portfolio elements.
  - ▶ Is the performance by one set (sample) of scorers generalizable to a different set of scores?
5. *Passing/Failing Decisions*—decision made based on the portfolio cut-score.
  - ▶ Is the decision made about a teacher based on his or her portfolio score consistent over time?

**Q16: Is there a finale to this chapter?**

**A16:** Kinda. After all the pages we've been through together, I'm beginning to have separation anxiety. The preceding review of research and practice with portfolios suggests that the teaching portfolio used in the context of licensure and certification decisions is a viable assessment tool to measure performance outcomes not measurable by alternative methods. Although the technical issues related to scoring, standard-setting, validity, and reliability are complex, they are not intractable. The portfolio requires more tender loving care than



constructed-response formats, but it provides valuable information on teacher performance. The track record and research of the CSBE and NBPTS furnish a foundation upon which other state departments can build. Since the portfolio can supply a more complete picture of teacher performance than many extant measures, it certainly deserves serious consideration.

## References

- Adkins, D. (1947). *Construction and analysis of achievement tests*. Washington, D.C.: U.S. Government Printing Office.
- Al Shehri, A. (1995). Learning by reflection in general practice: A study report. *Education for General Practice*, 7, 237–245.
- Arter, J. A. (1999). Teaching about performance assessment. *Educational Measurement: Issues and Practice*, 18(2), 30–44.
- Arter, J. A., & Spandel, V. (1992). Using portfolios of student work in instruction and assessment. *Educational Measurement: Issues and Practice*, 11(1), 36–44.
- Asher, J. J., & Sciarrino, J. A. (1974). Realistic work sample tests: A review. *Personnel Psychology*, 27, 519–533.
- Bashook, P. G., & Parboosingh, J. (1998). Recertification and the maintenance of competence. *British Medical Journal*, 316.
- Berk, R. A. (1999). Assessment for measuring professional performance. In D. P. Ely., L. E. Odenthal, & T. J. Plomp (Eds.), *Educational science and technology: Perspectives for the future* (pp. 29–48). Enschede, The Netherlands: Twente University Press.
- Brown, R. A. (1995). Portfolio development and profiling for nurses. *Journal of Clinical Nursing*, 4, 249–255.
- Cox, M. D. (1995). A department-based approach to developing teaching portfolios: Perspectives for faculty and development chairs. *Journal on Excellence in College Teaching*, 6(1), 117–143.
- Davis, M. H., Friedman Ben-David, M., Harden, R. M., Howie, P., Ker, J., McGhee, C., Pippard, M. J., & Snadden, D. (2001). Portfolio assessment in medical students' final examinations. *Medical Teacher*, 23, 357–366.
- Educational Testing Service. (1999, July). *National Board for Professional Standards: Technical analyses report*. Princeton, NJ: Author.



- Friedman Ben-David, M., Davis, M. H., Harden, R. M., Howie, P. W., Ker, J., & Pippard, M. J. (2001). Portfolios as a method of student assessment. *AMEE Medical Education Guide No. 24*.
- 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.)
- Gilman, D. A., & McDermott, M. (1994). Portfolio collections: An alternative to testing. *Contemporary Education, 65*, 73-76.
- Hall, B. W., & Hewitt-Gervais, C. M. (2000). The application of student portfolios in primary-intermediate and self-contained-multiage team classroom environments: Implications for instruction, learning, and assessment. *Applied Measurement in Education, 13*(2), 209-228.
- 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.
- Jasper, A. U. (1995). The potential of the professional portfolio for nursing. *Journal of Clinical Nursing, 4*, 249-255.
- Knapper, C. K. (1995). The origins of teaching portfolios. *Journal on Excellence in College Teaching, 6*(1), 45-56.
- Koretz, D. (1998). Large scale portfolio assessments in the US: Evidence pertaining to the quality of measurement. *Assessment in Education, 5*(3), 309-334.
- 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.
- 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.
- Murray, J. P. (1995). The teaching portfolio: A tool for department chairpersons to create a climate for teaching excellence. *Innovative Higher Education, 19*(3), 163-175.
- Seldin, P., Annis, L., & Zubizarreta, J. (1995). Answers to common questions about the teaching portfolio. *Journal on Excellence in College Teaching, 6*(1), 57-64.





- Shapley, K. S., & Bush, M. J. (1999). Developing a valid and reliable portfolio assessment in the primary grades: Building on practical experience. *Applied Measurement in Education, 12*(2), 111–132.
- Siegel, A. I. (1986). Performance tests. In R. A. Berk (Ed.), *Performance assessment: Methods and applications* (pp. 121–142). Baltimore, MD: Johns Hopkins University Press.

