DOCUMENT RESUME

ED 334 250 TM 016 834

AUTHOR Paulson, F. Leon; Paulson, Pearl R.
TITLE The Ins and Outs of Using Portfolios To Assess Performance. Revised.
PUB DATE May 91
NOTE 12p.; Expanded version of a paper presented at the Joint Annual Meeting of the National Council of Measurement in Education and the National Association of Test Directors (Chicago, IL, April 4-6, 1991).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Academic Achievement; Educational Assessment; Elementary Secondary Education; *Evaluation Methods; Evaluation Utilization; *Learning Processes; *Portfolios (Background Materials); *Student Evaluation
IDENTIFIERS Comparative Method; Environmental Beauty Estimation Method; Large Scale Programs; *Performance Based Evaluation; Stakeholder Evaluation

ABSTRACT

Concerns about using portfolios (collections of student work showing student effort, progress, or achievement in one or more areas) in large-scale assessments are addressed. The products in a portfolio allow the reviewer to make inferences about the process of student learning. Hence, a portfolio should include information about the activities that produced the portfolio and a narrative in which the student describes the learning that took place. Stakeholders in the portfolio review process are identified, and the role of instructional goals and determination of contents of the portfolio are discussed. Standardized input-output assessments that evaluators usually use are viewed as poorly suited to portfolios. The implications of chaos theory for educational measurement and, more specifically, portfolio evaluation is outlined. The use of multiple perspectives and differing criteria in analyzing portfolios are illustrated via a comparison with movie reviewers. The place of reliability and validity assessments in portfolio assessments and the use of generalizability theory are discussed. Two methods that accommodate the diversity required of portfolio assessments are outlined: (1) the Environmental Beauty Estimation Method used by the United States Forest Service, and (2) the Comparative Method used by sociologists in studying comparative political systems. The use of scaling techniques and the importance of holistic as well as analytic judgments are discussed. A 39-item list of references is included. (TJH)

The Ins and Outs of Using Portfolios To Assess Performance

Revised: May 1991

F. Leon Paulson
Multnomah Education Service District, Portland, Oregon

Pearl R. Paulson
Beaverton School District, Beaverton Oregon

A PORTFOLIO is a carefully crafted portrait of what a student knows or can do. It becomes a focal point for the student, teacher, parent, outside evaluator, and others. It is simultaneously a personal and a public statement. By *portfolio* we mean *a purposeful, integrated collection of student work showing student effort, progress, or achievement in one or more areas. The collection is guided by performance standards and includes evidence of student self-reflection and participation in setting the focus, selecting contents, and judging merit. A portfolio communicates what is learned and why it is important.*[1]

Our central concern is with the role of the student as portfolio owner, creator, and reviewer. Through building a portfolio, students have the opportunity to learn -- to learn about a subject, to learn about learning, to learn to make choices and judgments, and to learn about themselves. To us, the key issue is the process involved in creating a portfolio, not the products found in the portfolio.

This paper addresses concerns about using portfolios in large-scale assessments. We argue that the standardized input-output assessment model that evaluators usually employ is poorly suited, and that attempts to impose that model can do more harm than good. Before we develop this theme, let us describe what we mean by portfolio.

Portfolios: An Overview[2]

> I have in my possession photostatic copies of several pages of Beethoven's sketches for the last movement of his "Hammerklavier Sonata"; the sketches show him carefully modeling, then testing in systematic and apparently cold-blooded fashion, the theme of the fugue....The inspiration takes the form, however, not of a sudden flash of music, but of a clearly-envisioned impulse toward a certain goal for which the composer was obliged to strive.
>
> - Roger Sessions, composer

MENTION of the word *portfolio* often triggers a discussion about what should be placed *in* a portfolio. The assumption seems to be that in order to interpret what comes *out*, one needs to standardize what goes in. To us, the portfolio represents much more than the products placed into it. Rather, the products in a portfolio allow us to make inferences about process. Things find their way into the portfolio because students and teachers, working together, decide to put them there. The process of putting

Authors' Note: This is an expanded version of a paper read at a symposium conducted by the National Council on Measurement in Education and the National Association of Test Directors at their annual meeting in Chicago. We would like to thank Loni Myers (Beaverton School District, OR) and Sue Swanson (Mt. Scott School, Lake Oswego, OR) for providing the quotations from student portfolios appearing in this article.

1. This is is based on the the definition developed by the Northwest Evaluation Association (see Paulson, Paulson, & Meyer, 1991).

2. This is an overview of our Cognitive Model for Assessing Portfolios (CMAP). For a fully developed discussion, see F. L. Paulson & P. R. Paulson, 1990, 1991; P.R. Paulson & F. L. Paulson, 1991, in press.

things into the portfolio is far more important than the things themselves. This is the reason we emphasize that the portfolio collection include information about the activities that produced the portfolio (Paulson, Paulson & Meyer, 1991) and a narrative in which the students describe the learning that took place as they assembled their portfolios (P.R. Paulson & F.L. Paulson, 1991, in press). When we evaluate portfolios, we must find evaluation designs that protect this process, designs that do not impose external requirements for standardization.

Stakeholders

In defining the portfolio as a purposeful collection of student work, we must first ask *whose* purpose. A stakeholder is one who has a personal involvement or interest in the evaluation (see Guba & Lincoln, 1989; Stake, 1967). Clearly, this includes the student who may feel proud or vulnerable when someone reviews the portfolio. It also includes the teacher who may review the portfolio with satisfaction or disappointment, but may feel vulnerable when a supervisor reviews the same portfolio. Add the interests of parents, district evaluators, and even members of the school board, and the web of stakeholder interest becomes very complex indeed.

While there are many differences among portfolio stakeholders, the distinction between *primary* and *secondary* stakeholders is fundamental. The primary stakeholder is the individual who assembles, and therefore owns, the portfolio. Secondary stakeholders are all others who have some kind of interest in the portfolio. Certainly a portfolio developed by a student should address concerns held by the teacher who is, after all, the instructional leader. But the student as primary stakeholder has a personal stake in the portfolio that makes the portfolio unique.

Setting Goals

As evaluators, we tend to look at instructional goals from a top-down perspective. Instructional goals are set by curriculum committees and receive the breath of life from the classroom teacher. The portfolio requires us to rethink this model. How can we expect students to become self-directed learners if we insist that the goals worth assessing are those set by curriculum committees? The concept that the student is the primary stakeholder and the owner of the portfolio forces us to consider that the student will have goals as well. In fact, in a portfolio program, we expect students to begin setting personal goals and to assess progress toward their attainment.

There are two parts to goals setting; stating intentions and setting performance standards. Intentions establish the specific instructional focus, the outcomes or targets of an instructional program. Making students aware of program goals and involving them in setting, refining, and interpreting goals can have a positive effect on learning (Bereiter & Scardamalia, 1989), yielding one of the portfolio's major benefits.

Performance standards also have a role in learning. They frame expectations about what constitutes quality work (Wiggins, 1991). Standards are qualitative statements that describe outcomes in words so clear that students and other stakeholders can make judgments about the materials found in the portfolios. The descriptions (*not* the numerical scores) found in the direct writing assessment scoring rubric described by Spandel and Stiggins (1990) might serve as a prototype.

Contents

Stakeholders, especially the primary stakeholder, decide what goes in the portfolio. Stakeholders make these decisions in accordance with the stated rationale and issues following processes that may involve negotiation with other stakeholders. In service of student ownership of the portfolio, the interests of secondary stakeholders such as district evaluators must also be negotiated.

The contents of a portfolio can include a large variety of things; classroom assignments, finished or rough drafts, work students developed especially for their portfolios, self-reflections specific to issues, observations by teachers or other stakeholders, and so on. The potential for complexity requires that exhibit be organized and indexed in ways that provide a coherent picture.

Portfolios by nature become highly diverse, each reflecting the unique characteristics of the individual student.

Evaluation

Portfolios involve evaluation in a comprehensive sense. Students set the stage for evaluation when they collaborate with other stakeholders to describe the rationale, issues, and set standards. They develop their capacities to evaluate as they review and judge the quality of the work in their portfolios, an activity with profound implications over the long term.

Portfolios also create a context that requires the stakeholders to examine the portfolio in context and make informed judgments. Thus, portfolio assessment is more than data analysis, it is a process that involves disciplined inquiry in which the stakeholders review materials in context to make informed judgments. Stakeholders make inferences about the nature and quality of the learning that has taken place, both in specific areas of review or in judging the overall picture. The student as primary stakeholder has a major role in the activities that surround evaluation; and as students reflect on their learning and assess themselves as learners, they develop facility in using higher order thinking and metacognitive skills. Secondary stakeholders also evaluate the learning that is documented in the portfolio. Each stakeholder reviews the specific contents of the portfolio in relation to a personal set of intents and standards, and judges the portfolio according to a personal rationale and set of issues.

Stakeholders do not operate in isolation. They talk about what has been learned and why it is important. This communication among stakeholders is a most powerful contribution. It is the link between isolated activities in the classroom and the overall goals for an educational program.

Implications for Evaluators

> **What you've got to realise is that every cell in the nervous system is not just sitting there waiting to be told what to do. It's doing it the whole darn time. If there's input to the nervous system, fine. It will react to it. But the nervous system is primarily a device for generating action spontaneously. It's an ongoing affair. The biggest mistake that people make is in thinking of it as an input-output device.**
>
> - Graham Hoyle, neurobiologist

NOW that we have outlined our view of what portfolios might look like, we turn our attention to the issues that directly involve their assessment. One question, of course, is whether we should assess portfolios at all. Some argue that since portfolios play a major role in instruction, evaluators should not use them as sources of data for assessment. While there is an appeal to this argument, the toothpaste is already out of the tube. "Portfolio" appears with increasing frequency on lists of alternative assessment techniques used at district, state, and even national levels. Portfolio assessment at the state level is quickly becoming a reality (e.g., de Witt, 1991). Like it or not, portfolio *assessment* is here to stay. It is an engraved invitation to study and better understand complex mental processes, an assessment-rich environment that offers insight and understanding unavailable through more traditional methods. Portfolios have the potential to reveal a lot about their creators. They can become a window into the student's head, the means of understanding educational processes at a deeper level. They offer an appropriate means to assess what Resnick and Resnick (in press) call *the thinking curriculum.*

But as evaluators, we must approach portfolios with caution, avoiding the siege mentality that would turn this assessment-rich environment into a target-rich one. We must nurture the process, not undermine it. We must remain mindful of the nature of the thing we propose to study. Portfolios are neither standardized tests nor performance assessments although they provide a wonderful opportunity to assess students performing. Portfolios, however, provide highly authentic assessment. They are a natural environment, a cross section of student life that allows us to study the student in a relatively natural habitat. The challenge is to find appropriate evaluation and

measurement methods that allow access to that information.

We direct our comments at three concerns. One is that we reassess the way we think about reliability when we address portfolio assessment. The second is that we should seek analytic techniques that preserve the complexity. Third, we should be more restrained in our enthusiasm for scaling anything that moves.

Movie Reviews and Chaos: Rethinking Reliability

The way we usually think about reliability is based on the mathematics of test theory that assumes that the thing measured is linear and additive (Linn, 1984; Shepard, 1990). The problem is that the assumption is oversimplified. Human mental processes are not linear and additive and attempts to model human mental processes that assume linearity and additivity take us astray. Human behavior tends to follow unpredictable patterns; it is discontinuous and complex. To psychologist and artificial intelligence researcher Marvin Minsky (1986), the mind is a society of relatively independent operators that assemble and reassemble in ways that explain the apparent discontinuity in the ways we learn. Learning does not occur smoothly and in predictable increments. The neurobiologist William Calvin (1990) argues that the brain works like a group of self-organizing committees that take it in unpredictable directions. This capacity for unpredictability has evolutionary survival value (species with predictable behavioral patterns tend to become a meal for species with more flexible patterns). Psychologist and test theorist Lee Cronbach (1988) observes that human mental processes may be nonlinear, and may be described by the relationship found in the mathematics of chaos theory, models that are nonlinear and multiplicative.

Chaos.

Let us speculate on the implications of chaos theory for educational measurement. Chaos theory is a new way of thinking about natural events in which input-output determinism is replaced by the study of pattern. It employs nonlinear mathematics as a metaphor for the study of turbulence. Chaos was "discovered" in 1961 by Edward Lorenz at MIT when he was working with computer simulations of global weather patterns. Lorenz discovered that when he reran his computer models, very small changes in starting values produced sharp divergence in the weather patterns simulated (Gleick, 1987); minuscule changes in what went *in* produced huge changes in what came *out*. There was something provocative in the patterns Lorenz observed in his mathematically generated computer patterns. The patterns looked like real weather patterns.

There is much that is provocative in the patterns observed from a chaos theory perspective. Nonlinear patterns show up in very different and unexpected places. Medical researchers have found nonlinear analyses useful to describe and treat disorders of the heart. Biologists have begun to suspect that nonlinear processes may be a key to understanding how genes influence the growth of organisms. Ecologists are finding evidence of chaotic patterns in population growth, geographers in the shapes of shorelines, physicists in the variety of patterns found among snowflakes, economists in the ups and downs of the stockmarket (Gleick, 1978). Further, *both* theorists and experimentalists in widely separate disciplines observe these phenomena (Hofstadter, 1985). It is as if some of the best evidence for an orderly universe is found in events that have traditionally been thought chaotic.

Evidence has also begun to accumulate in the cognitive realm as well. The psychologists Cronbach and Snow (1977) analyzed hundreds of research and evaluation reports on the interaction of aptitude with treatment. They found thee interactions highly complex and difficult to generalize leading Cronbach (1975) to observe that looking at interactions was like entering a "hall of mirrors" that extends to infinity. Later (Cronbach, 1988) found an ideal simile in the language of chaos theory: "...like walking through a maze where walls rearrange themselves with every step you take" (p. 47, quoting Gleick, 1987 p. 24).

Gleick (1987) put chaos theory into a cognitive framework, noting that the fractal structure of chaotic models used in artificial intelligence research denotes a kind of infinitely

self-referential quality that is central to the mind's ability to produce ideas, make decisions, and experience emotions. Clearly, Chaos theory is opening doors to new ways to understand thinking. Hsu & Hsu (a father-son team comprising a geologist and a musician) demonstrated evidence that patterns found in fractal geometry (the analysis of chaotic patterns of self-similarity at all size scales which can be used to describe coastlines, snowflakes, and other natural phenomena) are present in things created by the human mind, for example, in the music of Bach (Browne, 1991). Arnold Mandel, a psychiatrist who uses chaos to study brain function, commented "when you reach equilibrium in biology, you're dead" (quoted by Gleick, 1987, p. 298), a statement that may apply to the mind as well as to the body.

Let us be clear about what we mean by "chaos." In popular usage, "chaos" denotes "complete disorder." Chaos *theory* brings a new meaning to the term: an addition to the dictionary. Chaos refers to deterministic patterns that are so complex that prediction is problematical. It is mathematical, but not statistical; probability plays no role. It is the uncanny resemblance of the chaotic patterns to patterns observed in nature that challenges the assumption that random processes are at work. The mathematical models used in educational measurement are statistical; probability plays a definite role. This suggests a problem. We may be using models that assume chaos in the popular sense (variation that is random) to measure processes that are chaotic in the technical sense (variation that is determined). If so, we may be discarding critical information as random error, a possibility that suggests we rethink our concept of error and the consequences of how we reduce it.

Movie Reviews.

Let us begin by looking at observer agreement. Complex, self-referential, chaotic systems like the human mind frequently disagree not only with one another, but with themselves. These disagreements are not random, they contain information. But in educational measurement we treat rater disagreement as error, random events with little or no informational value. If raters disagree, we conclude that our observations are unreliable and take steps to make them more "reliable" through procedures such as training. Siskel and Ebert illustrate our concern.

Gene Siskel and Roger Ebert are movie reviewers who present film reviews on a nationally syndicated television program. They are expert raters who decide whether or not to recommend particular films.[3] While their ratings are interesting, much of the most engaging information comes from their discussion, and, in particular, their disagreements. While Siskel and Ebert's disagreements obviously make good theater, their disagreements contain important information that might otherwise go unreported. Siskel and Ebert give information to clarify their differences, not resolve them. There is no pretense that there is one "right" answer to resolve to. (But, if their program were directed by an educational evaluator, Siskel and Ebert would probably be sent off to resolve their disagreements in order to present a united front on the air!)

Siskel and Ebert may disagree for several reasons. Occasionally their disagreements reveal different interpretations of the *same* criteria. For example, one criterion both use is empathy; do they care about the characters in a film. Disagreement does not signal the need for resolution. Rather, it provides information on how events can be viewed from multiple perspectives leading to differences in interpretation. They may attend to different information or weigh the same information differently. This suggests that when raters disagree on how to "score" something found in a student's portfolio, it may be more valuable to provide the student with a discussion of how and why the judges disagreed than to promote the illusion of a "united front" represented by a resolved score.

Occasionally, Siskel and Ebert disagree because they use *different* criteria, a possibility related to the stakeholder dimension of portfolio assessment. In portfolio assessment, students gain a valuable opportunity to learn from examining the criteria held by different stakeholders and by developing ways to accommodate to those divergent values and

3. They use a binary scale, "thumbs up" or "thumbs down," a topic we will return to later.

priorities. One student, struggling with this issue, wrote in his portfolio "I'm doing o.k. is what most people seem to think, but my mom says I'm doing it wrong. I don't know what to think." This student seems bewildered by the different criteria and is unable to use the information constructively. We as educators often recommend the use of one and only one set of criteria when judging student work, arguing that students will perform better when they know precisely what is expected. This, however, would not help the student in the example. This student's problem stems from the fact that the world outside the classroom applies different criteria from the world inside the classroom. Rather than working to develop common criteria to apply inside the classroom, our efforts might better be directed at helping students find ways to accommodate to the multiple criteria our pluralistic society routinely applies outside the classroom.

Reliability: Our servant, not our master.

Reliability is not a unitary concept where high is good, low bad. Reliability is a set of techniques we use to infer the degree to which we can place confidence in our observations and instruments. Achieving trustworthy observations is a standard, and reliability is a tool that can help us achieve that standard. As with any tool, we need constantly to re-examine its costs as well as its benefits, and, occasionally call for the fabrication of new tools and the development of new techniques.

Using instruments built on the statistical assumptions in test theory extracts a price and we must continually reassess that price. Defining rater disagreement as error may lead us to throw out extremely valuable information. Forcing agreement may obscure the fact that there are multiple criteria for judging performance and that perspectives differ with respect to how judgments are made. It is a practice that denies us information that might be valuable to students and other decision makers. Assuming we are measuring a process that is additive and linear when developing achievement tests may lead us to discard vital information as error. Using these tests may be like trying to listen to the Chicago Symphony on a crystal set. How we assess what is going on when we miss everything that occures outside an extremely narrow range?

Assessing portfolios requires that we seek models that refine and expand our ability to understand human performance as multidimensional phenomena occurring in complex, social contexts. Models that nudge us in that direction exist already. Generalizability theory, though linear, applies multifactor analysis of variance to test reliability (Cronbach, Gleser, Nanda, & Rajaratnam, 1972, Shavelson, Webb, & Rowley 1989). Generalizability theory is occasionally difficult to use, inconvenient to apply, and often produces puzzling results. But this may not be a shortcoming of the model. It may actually signal that the technique tunes into a world that is not as simple as classical test theory or item response theory would like to imagine (Shavelson, Carey & Webb, 1990), one in which simple, linear relationships are the exception rather than the rule. Suen and Davey (1990) have begun to address issues of reliability in their work with performance and portfolio assessment. In their view, reliability becomes less a goal to be attained than a tool to be manipulated.

As test developers, many of us think of high reliability as synonymous with "good" in some absolute sense -- a canon of our faith. We are quick to remind that one cannot achieve validity without first achieving reliability. The problem is when we set reliability as a *goal* in and of itself, then proceed to construct tests that produce high reliability coefficients. We employ mathematical models that assume unidimensionality, then we write unidimensional test items that produce results that conform to those models. We eschew "teaching to the test" by others while "testing to the test" ourselves. It recalls the old story of the person looking for a lost quarter under the street light because it too dark to look in the alley where the coin was dropped. Rather, let us explore the implications that the mathematical metaphors we routinely employ when assess reliability may be denying us access to important information.

Preserving Diversity

As we pointed out above. portfolios become highly diverse. In fact, we expect no two portfolios will be alike. But to some in the assessment community, this spells chaos.[4] "Whoa! Stop! It can't be done! How can you aggregate without standardization?" We argue that it can be done, although it may be less convenient to assess in the absence of standardized products. We argue against standardizing portfolio contents but encourage limited standardization of portfolio process (e.g., portfolios should have goals, performance standards, stakeholder input) although giving stakeholders wide latitude in their interpretation. When data are used for large scale assessments, we recommend using specific analytic techniques that accomodate diversity while ensuring rigor, impartiality, representativeness.

We will discuss two methods that work well in highly diverse environments. They are the *Environmental Beauty Estimation Method* that the U.S. Forest Service uses to make environmental management decisions, and the *Comparative Method* that sociologists use to study comparative political systems.

The *Environmental Beauty Estimation Method* (Daniel 1990; Daniel & Boster, 1976) is a scaling technique that satisfies the toughest requirements of reliability and validity in making aesthetic judgment across heterogeneous settings. It assumes that the construct, *scenic beauty*, is measured by the perceptual and judgmental process of humans when interacting with physical features of the environment. Second, it assumes that because the construct is not directly observable, it must be inferred. Finally, it assumes that the perceptual judgments of the general public provide an appropriate basis for judgment of the construct. Clearly, the Forest Service considers the public to be a stakeholder with respect to these judgments.

The approach uses classical psychophysical scaling techniques (Thurstone 1948; Stevens, 1958) that produce measures of scenic beauty on which observers agree, even highly diverse groups like professional foresters and environmental activists. The scale has been widely used in studies to quantify the impact of occurrences on environmental beauty, for example, the impact of logging on the desirability of recreation areas, the impact of air pollution on the scenic qualities of the Grand Canyon, or the impact of insect damage on the value of summer homes. The Forest Service does not require mother nature to have standardized trees in her portfolio (although reforestation seems to move in that direction).

Since we consider the portfolio an opportunity to study student performance in context, our second analytic technique is drawn from the world of ethnographic studies. While the use of quantitative data analysis techniques in sociology is widespread, sociologists also have techniques for use with descriptive and qualitative data. An advantage of qualitative techniques is that they are designed specifically to preserve diversity. This feature makes them attractive to researchers who worry that in many quantitative studies, the really interesting stuff ends up discarded as "error".

Ragin's (1987) *Comparative Method* employs boolean truth tables and boolean algebra. It uses binary classifications rather than scaled variables for analysis. The method proceeds to apply explicit, logical rules to reduce number of dimensions represented in the truth table without sacrificing the complexity represented in the classifications. The goal is to identify underlying aggregate clusters in the data. The analysis proceeds step-by-step to reduce the size of the table by making additional simplifying assumptions. The technique is similar to factor analysis with one important idfference. Simplifying assumptions are made as late as possible in an effort to preserve as much of the original information as possible.

The *Comparative Method* has been used in several settings (see Ragin, 1987). One study examined how descriptions of linguistically distinct ethnic populations of Western Europe and their degree of political mobilization. The challenge was to conduct comparative analyses across settings where there are no standardized features. The analysis uncovered ways qualitatively different combinations of descriptive factors, categorized in a binary way, were related to the similar outcomes. Two different combinations of descriptive factors (large size plus growing economic position, or

4. In the popular sense.

strong linguistic base plus high relative wealth) tended to produce a high degree of ethnic mobilization. Other combinations of the factors correlated with lower mobilization.

Scaling: When More Yields Less

Test developers usually seek ways to portray complex performance along a dimension that is *scaled* (using linear, additive assumptions) rather than *binary* (making no such assumptions). Many scoring rubrics used in performance testing clearly attempt to use scaling procedures. The analytic writing assessment scoring rubric described by Spandel & Stiggins (1990) and the holistic math scoring rubric from Project Equals (California Mathematics Council, 1989) are excellent examples of scoring rubrics that employ elements of scaling. Our concern, however, is that the evaluation community often assumes that a rubric with properties that can be scaled is *automatically and by definition* better than one that is not. We call this into question, based on our experience with the Oregon Preschool Test of Interpersonal Cooperation (The OPTIC system).

The OPTIC System (Paulson, 1976) is a situational response test (Weislogel & Schwartz, 1955), a forerunner of today's performance test. It was designed to test cooperative behavior in preschool children, thus providing a good example of an attempt to measure complex performance observed in context. Initially, we assumed that cooperative behavior was something that children learned in an incremental manner that could be scaled. Using literature searches and extensive observations of children interacting, we developed and tested several scoring rubrics. A typical 'scale' ran from (1) obstructive interaction, (2) minimal interaction, (3) active interaction, (4) pre-cooperation, and (5) full cooperation (Paulson, Paulson, Whittemore & McDonald, 1971; Paulson, 1972b).

Ultimately, we abandoned our attempts to scale cooperation for psychological as well as measurement reasons. Psychologically there simply was little evidence for an underlying, linear dimension of social behavior. Obstructive interactions were as likely to precede cooperation as pre-cooperative behavior. From a measurement point of view, our results were unstable over time and between social situations. When we adopted a binary scoring system, the quality of our results with respect to reliability (Paulson, 1972) and validity (Paulson, 1974) improved dramatically.

The Evaluator's Paradox

> **Zen koan:**
>
> **Shusan held out his short staff and said: "If you call this a short staff, you oppose its reality. If you do not call this a short staff, you ignore the fact. Now, what do you wish to call this?"**

WHAT do we evaluate when we evaluate? Do we evaluate the whole, or an accumulation of parts? Is there a difference? It recalls a Zen koan (above) that expresses we would call *the evaluator's paradox*. The paradox is that we need words to describe the staff, but we can't describe the staff with words. Words divide things into artificial categories and in so doing, the essence is lost. Evaluators need measurement facts to describe learning, but evaluators cannot describe learning using measurements facts. In our attempts, we deny its reality. Why? Probably because the facts give the appearance of capturing reality, yet they barely scratch the surface of reality. Categorization ignores an infinity of fact, the result is trivial (Hofstadter,1979).

The part vs. whole paradox is reflected in the debate between proponents of holistic and analytic approaches to writing assessment. The proponents of analytic assessment point out the diagnostic value of analytic information for the classroom teacher. But the value of analytic scores become is less clear when data are aggregated. How well does an aggregate score on an analytical trait, say sentence structure, answer the general question, "How well do students in Oregon write?" This is a holistic question that requires a holistic answer.

Direct writing assessment offers a prototype of assessment procedures that will probably be used with portfolios. Many are already choosing sides; analytic versus holistic. Our concern is that by joining either side, we lose. Here are three recommendations for assessing portfolios that, while they do not solve the part

vs. whole issue, are designed to keep the parts and the wholes in perspective.

1. Portfolio assessment should assess portfolios, not parts of portfolios. We often have good reason for looking at the parts of a portfolio, but we should always judge the parts in context. We may analyze samples which students wrote in, say, the persuasive mode, but when judging their ability to write persuasively, we should look at all entries in the portfolio for validation.
2. We should provide for *both* holistic and analytic judgment when designing portfolio coding systems and scoring rubrics[5], and use them in combination when judging. There is a great deal of information in portfolios. Let us judge that information with filters on and with filters off.
3. Let us recognize the evaluator's paradox and enjoy it for what it is -- a puzzle with many self-contradictory solutions. It makes our job as evaluators interesting, it keeps us honest, and it encourages humility. Unfortunately, it is a little difficult to explain to school boards.

Conclusion

THE thing portfolios do best is invite diversity. They give a perspective on student performance that is unique, pointing out that education is the product of many stakeholders, many points of view. Portfolios provide information on how pieces are integrated, looking more at process than product. They are, in a sense, a highly individual story (P.R. Paulson & F.L. Paulson, 1991, in press) of knowledge constructed by the learner, not supplied by the teacher. If done well, portfolio assessment has the capacity to reveal processes that are at the heart of learning. They allow us to adjust to the increasingly diverse populations of students coming through the doors of our schools.

When properly used in the classroom, portfolios become an invitation to think. They invite students to reflect on their learning, nurturing the ability to become independent, self-directed learners. But portfolios extend their invitation beyond the classroom. One teacher encouraged parents to write about things destined for their child's portfolio, an exercise that led a parent to write "It had the whole family thinking!" Evaluators are also on the guest list. We are challenged to think, reflecting on how we respond to this new opportunity. It requires us to move beyond the input-controlled world of the standardized test and to think about reliability, validity, scaling, and other measurement questions in new ways, ways that accomodate diversity in outputs. It calls on us to accept the challenge, adapting our methods to accomodate the needs of this new and somewhat enigmatic member of the classroom community.

References

Beaverton School District (1986). *Analytical trait writing: Student copy*. Beaverton OR: Beaverton School District 48J.

Bereiter, C. & Scardamalia, M. (1989). Intentional learning as a goal of instruction (pp. 361-392). In L. Resnick (Ed), *Knowing, learning, and instruction*. Hillsdale, NJ: Laurence Erlbaum Associates.

Browne, Malcolm W. (1991, April 16). In a musical invention, Bach + fractals = new compositions. *The New York Times*, pp. B5, B10. [Hsu, K. J. & Hsu, A. (1991) Self-similarity in the "1/f noise" called music. *The Proceedings of the National Academy of Sciences.*]

California Mathematics Council (1989). *Assessment alternatives in mathematics*. Berkeley CA: EQUALS, Lawrence Hall of Science.

Calvin, William H. (1990). *The cerebral symphony: Seashore reflections on the structure of consciousness*. New York: Bantam.

Cronbach, Lee J. (1975). Beyond the two disciplines of scientific psychology, *American Psychologist, 30*, 116-127.

5. These are different kinds of judgment. A total score that using weighted individual analytic trait scores does not yield holistic judgment.

Cronbach, Lee J. (1988). Playing with chaos [Review of Gleick's *Chaos: Making a new science*]. *Educational Researcher, 17(6)*. 46-49.

Cronbach, Lee J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability of scores and profiles*. New York: Wiley.

Cronbach, Lee J. & Snow, R. E. (1977). *Aptitudes and instructional methods: A handbook for research on aptitude-treatment interactions*. New York: Irvington.

Daniel, Terry C. (1990). Measuring the quality of the natural environment: A psychophysical approach. *American Psychologist, 45*, 633-637.

Daniel, Terry C. & Boster, Ron S. (1976). *Measuring landscape aesthetics: The scenic beauty estimation method* (USDA Forest Service Research Paper RM-167). Ft Collins CO: Rocky Mountain Forest and Range Experimentation Station. (Out of print)

de Witt, Karen (1991, April 24). Vermont guages learning by what's in a portfolio. *The New York Times*, p B7.

Gleick, James (1987). *Chaos: Making a new science*. New York: Viking.

Guba, E. G. & Lincoln, E. S. (1989). *Fourth generation evaluation*. Newbury Park, CA: Sage.

Hofstadter, Douglas R. (1979) *Godel, Escher, Bach: An eternal golden braid*. New York: Vintage.

Hofstadter, D. (1985). Mathematical chaos and strange attractors. In *Metamagical themas: Questing for the essence of mind and patterns* (pp. 364-395). New York: Basic Books. (Originally published in *Scientific American*, November 1981.)

Linn, R. L. (1984). Educational testing and assessment, *American Psychologist, 41*, 1153-1160.

Minski, Marvin (1986). *The society of mind*. New York: Simon and Shuster.

Paulson, F. L. (1972a). Live versus televised observations of social behavior in preschool children. *Proceedings, 80th Annual Convention of the American Psychological Association, 7*, 135-6. (ERIC Document Reproduction Service No. ED 071 735).

Paulson, F. L. (1972b). *The Oregon Preschool Test of Interpersonal Cooperation: Preliminary results*. Paper read to the Western Psychological Association, Portland.

Paulson, F. L. (1974). Teaching cooperation on television: An evaluation of "Sesame Street" social goals programs. *A.V. Communications Review, 22*, 229-246.

Paulson, F. L. (1976) The Oregon Preschool Test of Interpersonal Cooperation. In O. G. Johnson & J. W. Bommarito (Eds), *Tests and measurements in child development: Handbook II Vol.2*. San Francisco: Jossey-Bass, 1976.

Paulson, F. L. & Paulson, P. R. (1990) *How do portfolios measure up: A cognitive model for assessing portfolios*. Paper read at conference "Aggregating Portfolio Data" held by the Northwest Evaluation Association, Union WA, August. (ERIC Document Reproduction Service No. TM 015 516)

Paulson, F. L. & Paulson, P. R. (1991) *The making of a portfolio*. Unpublished manuscript available from the authors, 6800 Gable Parkway, Portland OR 97225.

Paulson, F. L., Paulson, P. R., & Meyer, C. A. (1991, February). What makes a portfolio a portfolio? *Educational Leadership, 46(5)*, 60-63.

Paulson, F. L., Paulson, P. R., Whittemore, S. L., & McDonald, D.L. (1971). *Handbook of information on interpersonal strategies in the behavior of young children*. Monmouth, OR: Teaching Research. (ERIC Document Reproduction Service No. Ed 057 895)

Paulson, P. R., & Paulson, F. L. (1991, in press) Portfolios: Stories of knowing. In P. H. Dreyer & M. Poplin (Eds), *Claremont reading conference 55th yearbook. Knowing: The power of stories*. Claremont, CA: Center for Developmental Studies of The Claremont Graduate School.

Ragin, C. C. (1987). *The comparative method: Moving beyond qualitative and quantitative strategies*. Berkeley: University of California Press.

Resnick, L. B. & Resnick, D. P. (in press). Assessing the thinking curriculum: New tools for educational reform. In B. R. Gifford & M. C. O'Connor (Eds), *Future assessments: Changing views of aptitude, achievement, and instruction*. Boston: Kluwer Academic Publishers.

Shavelson, R. J.. Carey, Neil B. & Webb, Noreen M. (1990). Indicators of science achievement: Options for a powerful policy instrument. *Phi Delta Kappan, 71*. 692 - 697.

Shavelson, Richard J., Webb, Noreen M., & Rowley, Glenn L. (1989). Generalizability theory. *American Psychologist, 44*, 922-932.

Shepard, L. A. (1990) Psychometricians's beliefs about learning. Paper presented at the annual meeting of the American Educational Research Association, April 17, 1990.

Spandel, V. & Stiggins, R. (1990). *Creating writers: Linking assessment and writing instruction*. New York: Longman.

Stake, Robert (1967). The countenance of educational evaluation. *Teacners College Record. 68(7).* 523-540.

Stevens, S. S. (1958). Problems and methods of psychophysics, *Psychological Bulletin. 55.* 177 - 196.

Suen, Hoi K. & Davey, Bruce (1990). Potential theoretical and practical pitfalls and cautions of the performance assessment design. Paper read at the annual meeting of the American Educational Research Association, Boston MA.

Thurstone, L.L. (1948) Psychophysical methods. In T. G. Andrews (Ed) *Methods in psychology*. New York: Wiley, 124 - 157.

Weislogel, R. L. & Schwartz, P. A. (1955). Some practical and theoretical problems in situational testing. *Educational and Psychological Measurement, 15*, 39-46.

Wiggins, Grant (1991, February). Standards, not standardization: Evoking quality student work. *Educational Leadership, 46(5)*, 18-25.