*Journal of Research Methodology, Volume 10, Number 1 (January - June 1997)*

*วารสารวิธีวิทยาการวิจัย ปีที่ 10 ฉบับที่ 1 (มกราคม-มิถุนายน 2540)*

# The Validity and Reliability of Information from a Student Annotation Form for Work Samples in Portfolios

Susan M. Brookhart*

**ABSTRACT**

*This paper presents evidence for the validity and reliability of information from a Student Annotation Form used to collect student reflections on work samples in portfolios for use in a districtwide evaluation. The forms were designed to be general enough to apply to many grades and subjects, so that they could be used across classes in the evaluation. The annotation forms were simple enough for students at all levels to complete. The forms asked students to indicate how difficult they found the work sample to be, whether they would like to do more of the work, and why. The difficulty and do-more questions were multiple choice; the "why" question was open-ended. For each work sample that illustrated a curriculum objective, selected for inclusion in an evaluation portfolio, students completed an annotation form and attached it to the work sample. Study One was performed with data collected in 1994, from a total of 367 students, grades 1 through 10, in whose portfolios were included a total of 1678 annotated work samples. The validity and reliability of the difficulty and do-more judgments were confirmed. A constructed measure of academic self-efficacy looked promising at first but failed a validity check. Study Two was performed on data collected in 1995, from a total of 313 students, grades 1 through 10, in whose portfolios were included a total of 2862 annotated work samples. The quantitative reliability and validity results were replicated for the difficulty and do-more questions.*

* Duquesne University

A version of this paper was presented at the 1996 Annual Meeting of the American Educational Research Association. This work was supported in part by a Duquesne University Presidential Summer Writing Scholarship, and in part by the Fox Chapel School District. The analysis and conclusions expressed here are the responsibility of the author. Thanks go to Laurel J. Swenson for her assistance with this study.

Various kinds of portfolios, constructed according to all sorts of different rules, are being used to assess student achievement, progress, attitudes, and understandings for many different purposes, including regular classroom assessment, evaluation of school programs, and evaluation of district-level achievement. One common element in all portfolios, an element that distinguishes them from traditional measures of achievement and, some say, is the key to their usefulness, is the opportunity they offer for students' reflections on their work.

Many forms, procedures and strategies have been suggested for student reflection on portfolio work. These include checklists of work done and reflection questions to answer (Farnan & Kelly, 1991), letter-writing or summarizing assignments (Ferguson & Maples, 1992), and various designs of self- evaluation forms, for annotating individual work samples or for reflecting on the whole portfolio (Azar et al., 1993; Goodman , Bird, & Goodman 1992). There is some evidence that teachers' use of student reflection increases as they work with portfolios over time (Roe & Vukelich, 1995). Reliability and validity studies have typically been performed on the ratings of achievement or performance in portfolios (e.g., Klein, McCaffrey, Stecher, & Koretz, 1995), however, not on student reflections. The results of these studies of the technical qualities of porfolio achievement measures have been mixed (Calfee & Perfumo, 1993; Herman & Winters, 1994). No studies were found on the technical qualities of portfolio measures of student reflections or dispositions toward their work.

This paper presents evidence for the validity and reliability of information from a student annotation form used to collect student reflections on work samples in porfolios for use in a districtwide evaluation. The forms were designed to be general enough to apply to many grades and subjects, so that they could be used across classes in an evaluation project. Part of a larger evaluation project for the Fox Chapel (PA ) Area School District required the design of achievement and student progress measures that tapped some outcomes of instruction not measured by standardized tests or by teacher judgment, since these data are readily available in other measures. Specifically, the measures were to tap student judgments about their academic work and themselves as academic workers. These kinds of judgments contribute to academic self-efficacy (Zimmerman & Martinez- Pons, 1992), and they are an important and poorly-measured aspect of what the general and political public at present calls "21st century skills." Beyond problem-solving ability per se, there are orientations that will genuinely establish today's students as tomorrow's problem-solvers. Not everyone who can solve a problem, does. The efficacy measures in this evaluation were to tap feelings of efficacy at doing the particular work

sampled in the portfolio, not the more general academic self-efficacy for which standardized measures are available

One of the measures designed was a Student Annotation Form, to be completed and attached to work samples in student portfolios. For each work sample that illustrated a curriculum objective, selected for inclusion in an evaluation portfolio, students completed an annotation form and attached it to the work sample. The design of the Student Annotation Form was influenced by previous experience in a pilot study. The pilot study had used a more open-ended annotation form that asked, " What did you learn?" The data gathered with this form had been more likely to include the title of a page of work ,e.g., "Two-digit multiplication," than anything reflective. When a student copied the assignment title, it was not certain that the student knew what it meant. Thus the Student Annotation Form (see Figure 1) asked students to make two multiple-choice judgments, about the difficulty of the work sample and whether or not they would like to do more of that kind of work, and to write briefly "Why" they would or would not like to do more.

**Figure 1** Student Annotation Form (Student copies were half-sheets to be stapled to work sample

**Circle your choices and then tell why, Staple this paper to your work sample**

**I found this work**
**a. easy**
**b. about the same as other work.**
**c. hard**

**I would like to do more of this kind of work**
**a. yes**
**b. no**

**Why?**

The expectation was that in responding to the "Why" question, students would tip their hands about what they thought the assignment meant, at least for themselves. The choices for student judgment of difficulty were based in common classroom talk. Many students say a task was "easy" when they mean to boast "I could do it, and I did it successfully." In the experience of the evaluation team who developed the Student Annotation Form, a student judgment that work was "easy" did not necessarily mean that it was beneath their capabilities, at the level of unchallenging busywork, but it was a judgment they made naturally. The opposite of "easy," for students, is "hard"

The use of multiple-choice questions to stimulate reflection is not recommended as a strategy for portfolios intended to be used primarily for instruction. The intended use of information here, however, was as an indicator of student perceptions of work done under a program called Continuous Progress Instruction, based on Wang's (1992) Adaptive Learning Environments Model. Most of the classes participating in the initial implementation phase were mathematics classes, but some were English classes. Student age and developmental level varied widely, so the multiple choice items were viewed as basic indicators for the evaluation that all students could do and that would lead some students to deeper reflection in the open-ended responses that followed. This study is concerned with validating the use of the student annotation form for this cross-classroom evaluation.

Student effort and achievement has been shown to be related to both perceptions of task characteristics and students' belief in their abilities to handle those task characteristics (Salomon, 1984). Therefore, the evaluators experimented with a measure of self-efficacy to do the particular kind of work demonstrated in the portfolio. The judgment of difficulty (DIFF=1 if the work was judged easy, 2 if about the same as other work, and 3 if hard) was multiplied by the willingness to do more of that particular kind of work (DOMORE=1 if yes , 0 if no). Thus the self-efficacy indicator (SE) ranged from 0 (would not like to do more) through 1 (would like to do more of this easy kind of work), 2 (would like to do more of this average kind of work), to 3 (would like to do more of this hard work). It was reasoned that work a student was unwilling to do would not contribute to a student's image of himself or herself as a worker in that area and that as students were willing to tackle work they judged progressively more difficult, they would have progressively stronger images of themselves as workers in that area. In addition to results for the validity of information from the two items on the student annotation form, this study presents results of a validity check on the constructed measure of self-efficacy. Briefly, the items on the form itself were found valid for their intended districtwide evaluation purpose, but the experimental constructed measure of work sample- dependent self-efficacy was not found to be valid.

## Research Question

The annotation forms designed for the Fox Chapel evaluation offer a very simple and widely applicable tool for student reflection. But before this tool is widely applied, it is very important to establish that the information elicited by these forms is valid and reliable, that is, that the responses are accurate and dependable representations of what the students think. The research question under investigation was as follows: What evidence is there for the validity and reliability of student annotations of work samples in portfolios?

Specific to validity, do student's ratings of the difficulty of the work match their perceived ability to learn the material? When a student reports wanting to do more of the work, what reasons are given? Are these reasons conducive to learning? Does the constructed self-efficacy variable (formed by multiplying the difficulty and do more responses) match students's perceived ability to learn the material?

Specific to reliability, what is the internal consistency of the set of annotations in each student's portfolio? What is the generalizability of responses across time (as the students use the same annotation form over and over)? Internal consistency and generalizability were important to investigate because in the evaluation studies, the annotations were aggregated across portfolios to arrive at student overall ratings of the difficulty of their work and their willingness to do more. These uses required that the annotations may be aggregated reliably. For using information from an individual annotation and its work sample in classroom instruction, the validity questions about the relationship between student reasons for ratings and learning, listed above, would be of primary importance.

Study One includes results of a complete investigation of these research questions using data collected in 1994. Reliability and validity of information from the DIFF, DOMORE, and SE variables were studied with both quantitative and qualitative methods. Qualitative analysis was required for the open-ended responses to the "Why?" question. Study Two includes results of a replication of the quantitative analyses of the DIFF and DOMORE variables, using data collected in 1995. Analyses of the SE variable were eliminated because of the results of Study One.

# Study One

## Method

***Data.*** Data were collected in 1994. Within each class in the evaluation project, portfolios from 10 students were selected according to a stratified (gifted, nonlabeled, and educational support) random sample. There were portfolio data from a total of 367 students, grades 1 through 6 and grade 10, in whose portfolios were included a total of 1678 annotated work samples. There were students from every school in the district: 4 elementary schools, one middle school (the 6th graders) and the high school. Most of the portfolios were from mathematics classes (82%); the rest were from English/Language Arts classes (18%). Fifty-four percent of the students were male; 46% were female.

Student status in this sample was as follows: gifted, 23%; nonlabeled, 58%; educational support, 19%. All classes in the sample were part of the initial districtwide implementation of Continuous Progress Instruction.

Data included student responses to each Student Annotation Form (number of annotated work samples per student portfolio ranged from 0 to 13 ) in the portfolio and teacher ratings from a cover sheet, preprinted with appropriate curriculum objectives for the class, that the teachers inserted in each student's portfolio. Teacher ratings of achievement used a 1-4 rubric: The student demonstrates (1) no evidence of mastery of the objective, (2) partial/incomplete mastery, (3) acceptable mastery, or (4) exceptional mastery. Teachers rated efficiency of use of study time for each curriculum objective demonstrated in the portfolio (1=efficient, 0=not efficient). Background information (grade, subject, school, student status, and gender) about students was included in the data set. Mean DIFF, DOMORE, and SE ratings were calculated for each student's portfolio, based on the total number of annotations in that portfolio. Missing data resulted in varying sample sizes for each analysis; actual n for each analysis is reported in the tables.

***Analyses.*** Two reliability analyses were done. (1) Cronbach's alpha was calculated for each rating, treating annotations within one student's portfolio as items on DIFF, DOMORE, and SE scales, respectively. Student was the unit of analysis. Internal consistency reliability is required if meaningful composite DIFF and DOMORE ratings for each student are to be calculated. Overall mean DIFF and DOMORE ratings were used in the evaluation, to answer evaluation questions about students' reported dispositions toward their work.

(2) Generalizability studies were conducted for each of the three scales, with the design Person by Time. The Time facet represented the ordinal position of the annotation in the student's portfolio: the first time the student used the form, the second time, and so on. This design examined both generalizability across time and the size of the Person-by-Time interaction. Generalizability studies were conducted to answer two reliability questions. First, how many annotations per portfolio are required for a reliable estimate of students' overall dispositions toward their work? Second, does using the same form over and over result in patterns of responses that might indicate students tire of reporting DIFF and DOMORE? If so, questions are raised about the validity of continued use.

Additional quantitative and qualitative validity analyses were conducted. Correlations of DIFF, DOMORE, and SE with overall teacher ratings of achievement, effectiveness of use of study time, and percent of curriculum objectives attained were examined.

Mean ratings of achievement and effectiveness of time use for each student were calculated over all curriculum objectives that the student attempted; number and level of objectives attempted differed for each student. Percent of objectives attained was calculated as the number of objectives on which a student received as satisfactory or better teacher rating, divided by the total number of objectives marked on the student's porfolio cover sheet. The student was the unit of analysis. These correlations provided a look at the external structure of information from the Student Annotation Forms. The external measures were teacher ratings of achievement and time use on the same curriculum objectives for which work samples were selected and annotated for the portfolios.

Two qualitative analyses of the open-ended responses addressed these research questions: Do students' ratings of the difficulty of the work match perceived ability to learn the material? When a student reports wanting to do more of the work, what reasons are given? Are these reasons conducive to learning? What reasons for wanting to do more of the work are given at each level of self-efficacy, and are these reasons consonant with the interpretation of academic self-efficacy the experimental variable is thought to measure? Individual annotations were the unit of analysis. First, student responses to the open-ended question "Why?" were coded according to how the student viewed further learning; (a) response contained reasons or opinions that would facilitate further learning for the student, (b) reponse contained reasons or opinions that would hinder further learning, (c) neutral or ambiguous, and (d) response contained reasons or opinions that would both facilitate and hinder further learning for the student. To check the reliability of coding, two independent coders categorized a random sample of 100 of the written comments into these four categories; their agreement rate was 87%. Table 1.1 illustrates these codes with some example student responses from each category.

The second analysis examined key words in the text of students' written annotations. The words "easy," "fun," and "hard" occurred often. These words are also commonly used in students' talk about schoolwork. Word searches on each of these words were done by computer, then augmented by visual inspection of the text. Inspection allowed for adding misspellings and invented spellings that the computer did not find, for example, "ese" for "easy" or"hrd" for "hard," and to add variations that were not found in the computer search, for example, "easier." Inspection also allowed the separation of "too easy," a negative response, form "easy," which students usually use as a positive description of schoolwork. If a word was repeated within one annotation, it was counted only once for example, the comment because it is easy and I like easy work " was counted as one expression of "easy." Responses were organized by DIFF, DOMORE, and SE values, then a table was created tallying word use at each rating level.

**Table 1.1** Examples of Student Written Comments about Why They Would or Would not Like to Do More of the Type of Work Represented by a Work Sample

| Sample Comment | Learning Code |
|---|---|
| Because I got a better understanding of the book and I saw what other people thought of the book. | 1, facilitates learning |
| Because using division patternes are easy and hard, and if I do more I could probably get a A on my test. | 1, facilitates learning |
| Because I don't like word problems | 2, hinders learning |
| I thank it is the wors papr in the world and I do not want to do it agane !!!!! | 2, hinders learning |
| A different y way to learn | 3, neutral or ambiguous |
| I guess that this problem is ok, I still think it should explain more. I just don't like the extra enrichment. It takes to long. But it lets us go at our own rate. | 4, both facilitates & hinders |

## Results and Discussion

***Reliability.*** In general, reliability was acceptable for the districtwide evaluation purpose. Alpha reliability for seven annotation forms per portfolio was .62 for difficulty ratings, .71 for the domore ratings, and .69 for academic self-efficacy. Table 1.2 presents alpha values for 6 through 9 annotations per portfolio. Very few of the portfolios included more than 9 annotated work samples. Internal consistency of annotation of work samples within a portfolio would clearly depend in part on the similarity of the work samples to one another; it was not expected that the annotations would be as internally consistent as the items on a test or survey.

**Table 1.2** Reliability (Internal Consistency) of Scales, for Six through Nine Annotations per Portfolio

| Scale | k | $\alpha$ | n | $\alpha'$ |
|---|---|---|---|---|
| DIFF | 6 | .59 | 158 | .71 |
| | 7 | .62 | 121 | .70 |
| | 8 | .60 | 53 | .65 |
| | 9 | .64 | 20 | .66 |
| DOMORE | 6 | .66 | 157 | .76 |
| | 7 | .71 | 118 | .78 |
| | 8 | .73 | 53 | .77 |
| | 9 | .80 | 19 | .82 |
| SE | 6 | .65 | 153 | .76 |
| | 7 | .69 | 116 | .76 |
| | 8 | .74 | 51 | .78 |
| | 9 | .84 | 19 | .85 |

$\alpha'$ = Reliability estimate for k =10, using the Spearman-Brown formula

If reliability is considered to be the usefulness of the annotations across time, generalizability is at issue.As students use the same annotation form over and over again, for successive work samples included in their portfolios , do their judgments waver? For the DOMORE and SE variables , in fact , the variance due to the Time factor was zero (see Table 1.3), leading to identical values for both absolute and relative generalizability in the Person by Time design studied. The Person-by-Time interaction for each was large (about 80% of total variance). This means that different students did change the difficulty and do-more ratings in somewhat different ways over time; in part, this reflects the fact that work samples themselves differed. The large Person-by-Time interaction also showed that the student annotation forms collected data that behaved in a similar fashion to most performance assessment data, which typically has a large person-by-task variance component (Brennan , Gao,& Colton, 1995). So while one of the challenges of instrument development will be to reduce this term, at present its size suggests that students approach the student annotation forms in a way more similar to how they would approach performance assessments than paper-and-pencil surveys. All three scales generalized at an acceptable level across uses (see Table 1.4). Generalizability coefficients for these three ratings were similar.

**Table 1.3** Variance Component Estimates for Scales, for Person x Time Design

| Source of Variability | DIFF Estimated Variance Component | DIFF Percent Total Variability | DOMORE Estimated Variance Component | DOMORE Percent Total Variability | SE Estimated Variance Component | SE Percent Total Variability |
|---|---|---|---|---|---|---|
| Person | .0913 | 19% | .0527 | 26% | .1640 | 24% |
| Time | .0036 | 1% | .0000 | 0% | .0000 | 0% |
| P x T | .3923 | 81% | .1495 | 74% | .5242 | 76% |

DIFF & DOMORE variance estimates based on 118 persons, 7 times
SE variance estimates based on 116 persons, 7 times

**Table 1.4** Generalizability of Scales, for Six through Ten Annotations per Portfolio

| Scale | k | $\hat{\rho}^2$ | $\hat{\phi}$ |
|---|---|---|---|
| DIFF | 6 | .58 | .58 |
| | 7 | .62 | .62 |
| | 8 | .65 | .65 |
| | 9 | .68 | .67 |
| | 10 | .70 | .70 |
| DOMORE | 6 | .68 | .68 |
| | 7 | .71 | .71 |
| | 8 | .74 | .74 |
| | 9 | .76 | .76 |
| | 10 | .78 | .78 |
| SE | 6 | .65 | .65 |
| | 7 | .69 | .69 |
| | 8 | .71 | .71 |
| | 9 | .74 | .74 |
| | 10 | .76 | .76 |

DIFF & DOMORE generalizability estimates based on 118 persons
SE generalizability estimates based on 116 persons

***Validity.*** As the rationale for these annotation forms suggests, it was expected that these forms would tap self-efficacy about particular schoolwork by indicating student judgments about the difficulty of the work and their willingness to pursue it. Table 1.5 shows there were weak but significant correlations between students' mean DIFF ratings and teacher judgment of how efficiently overall students used their time in independent work. (r=-.17, p=.005) and between students' mean DOMORE ratings and teacher judgment of how efficiently students used their time (r=.12, p=.05). Teachers judged more efficient time use for students who rated their work samples less difficult overall. Teachers judged more efficient time use for students who indicated more interest in doing their work. This weak convergent evidence suggested that students' in-class behavior was somewhat consistent with their annotations.

**Table 1.5** Correlations between Scales and Selected External Measures

| | PCTACH | TIME | EVAL |
|---|---|---|---|
| SE | -.02 | -.01 | .04 |
| DIFF | -.09 | -.17** | -.10 |
| DOMORE | -.02 | .12* | .03 |

*p=.05, **p=.005, n ranged from 269 to 286 students
PCTACH = percent of learning objectives attempted within the portfolio that were achieved
TIME= efficiency of time use, 0=inefficient, 1=efficient, averaged over all learning objectives represented in the portfolio
EVAL= teacher evaluation of achievement, on a scale of 1 to 4, averaged over all learning objectives represented in the portfolio

None of the three scales (DIFF, DOMORE, or SE) was significantly correlated with the percent of attempted learning objectives, documented in the portfolio, that were achieved (each portfolio was supposed to include one work sample per learning objective), as measured by teacher judgment, or with teacher judgment of the level of achievement indicated in the portfolio. This divergent evidence suggested that the student annotation forms were not measuring the same thing as teacher judgment of achievement

Cross- tabulations between DIFF, DOMORE, and SE scales (Tables 1.6, 1.7, and 1.8, respectively) and the four LEARNING categories showed expected relationships. For example , for annotations on which the students responded yes, they would like to do more of the kind of work in the sample, 91% of the written responses were about the work facilitating further learning; conversely, for annotations on which the students responded no, they would not like to do more of that kind of work, 73% of the written responses gave evidence of hindering further learning (Table 1.7)

**Table 1.6** Crosstabulation of Students' Judgments of Difficulty with Their Judgments about the Value of the Sampled Work for Their Learning

| LEARNING | DIFF | | | |
|---|---|---|---|---|
| Frequency (Col Pct) | EASY | SAME AS OTHER WORK | HARD | Total |
| FACILITATES | 674 (77%) | 326 (64%) | 68 (33%) | 1068 (67%) |
| HINDERS | 134 (15%) | 125 (25%) | 118 (57%) | 377 (24%) |
| NEUTRAL/ AMBIGUOUS | 36 (4%) | 27 (5%) | 13 (6%) | 76 (5%) |
| BOTH | 36 (4%) | 29 (6%) | 9 (4%) | 74 (5%) |
| Total | 880 | 507 | 208 | 1595 |

Unit of analysis = annotation of one work sample
LEARNING was code from student responses to open-ended question, "Why?"

**Table 1.7** Crosstabulation of Students' Willingness to Do More with Their Judgments about the Value of the Sampled Work for Their Learning

| LEARNING Frequency (Col Pct) | DO MORE NO | DO MORE YES | Total |
|---|---|---|---|
| FACILITATES | 67 (14%) | 1003 (91%) | 1070 (67%) |
| HINDERS | 356 (73%) | 26 (2%) | 382 (24%) |
| NEUTRAL/ AMBIGUOUS | 26 (5%) | 48 (4%) | 74 (5%) |
| BOTH | 40 (8%) | 28 (3%) | 68 (4%) |
| Total | 489 | 1105 | 1594 |

Unit of analysis = annotation of one work sample
LEARNING was coded from student responses to open-ended question, "Why?"

**Table 1.8** Crosstabulation of Student's Self-Efficacy about the Work with Their Judgments about the Value of the Sampled Work for Their Learning

| LEARNING Frequency Col Pct | SELF-EFFICACY 0 | 1 | 2 | 3 | Total |
|---|---|---|---|---|---|
| FACILITATES | 67 (14%) | 630 (93%) | 306 (89%) | 61 (78%) | 1064 (67%) |
| HINDERS | 356 (73%) | 9 (1%) | 8 (2%) | 8 (10%) | 381 (24%) |
| NEUTRAL/ AMBIGUOUS | 26 (5%) | 25 (4%) | 15 (4%) | 7 (9%) | 73 (5%) |
| BOTH | 40 (8%) | 12 (2%) | 14 (4%) | 2 (3%) | 68 (4%) |
| Total | 489 | 676 | 343 | 78 | 1586 |

Unit of analysis = annotation of one work sample
SE=DIFF X DOMORE
LEARNING was coded from student responses to open-ended question, "Why?"

Correlations of mean DIFF, DOMORE, and SE for each student with the mean LEARNING indication in a student's written comments (transformed into 1=facilitates, 0=hinders, disregarding ambiguous responses) indicated that self-efficacy for learning and wanting to do more of a particular kind of work were positively related to perceptions of usefulness for learning, while perceived difficulty level was negatively related to perceptions of usefulness for learning. This is consistent with cognitive theory that says students need a moderate level of challenge. These relationships held, in the same strength and direction, for gifted, nonlabeled, and educational support students, with the exception of the relationship of DIFF to LEARNING for gifted students (see Table 1.9). These relationships also held for the most part across grade levels; the exceptions were grades 2 and 5 (see Table 1.10).

Usage of key words from the students' own writing are organized in Table 1.11 according to the DIFF, DOMORE, and SE values for the annotation on which the comment appeared. The frequencies and percents in the table may be considered a conservative estimate of usage. To keep the search from requiring much inference on the part of the reader, only the key words were counted. Invented spellings were allowed, but possible conceptual connections were not. For example, one student wrote "I already no [know] it" and might well have considered the work easy or too easy, but the comment was not tallied as an instance of "easy." The only conceptual inference permitted in the search was the removal of negatives as an instance of usage. For example, one student wrote,"I'd rather do something fun, "and this was not counted as an instance of fun, since the student clearly did not think this work was fun.

**Table 1.9** Correlation between DIFF,DOMORE, and SE Scales and Students' Judgments of the Value of the Work to Their Learning, Overall and by Student Status

| | | Status | | |
|---|---|---|---|---|
| | All Students | Gifted | Nonlabeled | Educational Support |
| n | 308 | 68 | 185 | 55 |
| DIFF | -.22 | .02 | -.28 | -.36 |
| DOMORE | .80 | .81 | .76 | .93 |
| SE | .63 | .73 | .57 | .74 |

Unit of analysis = student
Learning judgment is the average, of all annotations in a student's portfolio, of written responses to "Why" s/he would do more of the kind of work in the sample, coded 1 = facilitates learning, 0= hinders learning (ambiguous responses were not used in this analysis.

**Table 1.10** Correlation between DIFF, DOMORE, and SE Scales and Students' Judgments of the Value of the Work to Their Learning, Overall and by Grade

| | | Grade | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | All Students | 1 | 2 | 3 | 4 | 5 | 6 | 10 |
| n | 308 | 49 | 52 | 55 | 27 | 59 | 21 | 45 |
| DIFF | -.22 | -. 34 | .21 | -.55 | -.12 | .01 | -.28 | -.32 |
| DOMORE | .80 | .80 | .93 | .95 | .63 | .64 | .75 | .85 |
| SE | .63 | .59 | .80 | .61 | .48 | .56 | .67 | .73 |

Unit of analysis = student
Learning judgment is the average, of all annotations in a student's portfolio, of written responses to "Why" s/he would do more of the kind of work in the sample, coded 1 = facilitates learning, 0= hinders learning (ambiguous responses were not used in this analysis.

**Table 1.11** Content Analysis of Frequently Used Words from "Why?" Responses on Student Annotation Forms

| | | Word(s) | | | |
|---|---|---|---|---|---|
| Ratings | Total | Too easy | Easy | Fun | Hard |
| SE = 0 | 482 | 58 (12%) | 56 (12%) | 4 (1%) | 93 (19%) |
| DIFF = 1, DOMORE = 0 | 199 | 53 (27%) | 34 (17%) | 3 (2%) | 10 (5%) |
| DIFF = 2, DOMORE = 0 | 157 | 4 (3%) | 17 (11%) | 1 (0%) | 19 (12%) |
| DIFF = 3, DOMORE = 0 | 126 | 1 (0%) | 5 (4%) | 0 (0%) | 64 (51%) |
| SE = 1<br>DIFF = 1, DOMORE =1 | 677 | 1 (0%) | 238 (35%) | 194 (29%) | 8 (1%) |
| SE = 2<br>DIFF = 2, DOMORE =1 | 341 | 0 (0%) | 41 (12%) | 93 (27%) | 15 (4%) |
| SE = 3<br>DIFF = 3, DOMORE =1 | 77 | 0 (0%) | 3 (4%) | 12 (16%) | 16 (21%) |

Values in the table are frequencies of occurrence of each word, and percent of total usage for each rating level of the work

The information in Table 1.11 suggested that usage of "easy" and "hard" varied as expected with DIFF ratings. This supported the validity of these ratings. Observations about key word usage by SE indicated that for SE=0 (and thus for DOMORE=0), "fun" was conspicuously absent. For SE =1 and SE =2, "easy" and "fun" work was more common. For SE =3, "hard" work was more common.

Visual inspection of the comments suggested that the tone of the instances of "hard" for SE=0 was very different from the instances for SE=3. It appeared that the "hard" work for SE=3 was seen as a challenge, while for SE=0 hard work was a stumbling block. The 126 annotations for which DIFF=3, DOMORE=0 and the 77 annotations for which DIFF=3, DOMORE=1 were reread, for themes, and although this process did require making some inferences, the results were very clear.

Twenty-three of the 77(30%) annotations where the student indicated he or she wanted to do more hard work included some comment about learning more, getting better with practice, or value for future work. Twenty-four of the 77(31%) included comments about enjoying the work. Fourteen of the 77 (18%) included statements about challenge. None of the 77 annotations for DIFF= 3, DOMORE=1 indicated that the work was too hard.

Conversely, 8 of the 126 (6%) annotations where the student indicated he or she did not want to do more of the hard work included the words "too hard." Thirty-one of the 126 (25%) annotations simply pronounced the work hard, with no explanation. Twenty-two of the 126(17%) annotations included a statement about not enjoying the work. Only one (1%) mentioned a challenge.

The validity evidence from qualitative analyses of the students's own writings was very compelling, with one exception. Learning attitudes demonstrated by the annotations were consistent with the DIFF, DOMORE, and SE ratings. Key words and concepts were consistent with these ratings, too, with the exception of the 53 annotations whose student comments read they did not want to do more easy work because the work was "too easy", plus a few with similar viewpoints that did not use the phrase "too easy." The tone of these comments was confident and did not evince lack of self-efficacy for learning.

Thus the constructed measure of students' self-efficacy to accomplish their work did not stand up to a validity study. Data from the SE variable (DIFF X DOMORE) stood up to reliability checks and preliminary validity investigations. However, a content analysis of comments students wrote indicated that at least some of those students who did not want to do more of the kind of work on the work sample, even though they reported it was easy

work, felt that way because they considered the work too easy and therefore mastered. Wanting to move on to harder work because easier work is mastered is not consistent with a score of "0" on the SE scale.

Since the other categories of the SE scale did seem to match with student comments, especially noticeable in the comments of SE=3 students who reported enjoying handling academic challenge, the search for an SE measure from these annotations should not be abandoned. Portfolios present an opportunity for students to reflect on the achievement represented by particular work samples. The potential exists for a more specific measure of self- efficacy than heretofore available.

## Study Two

## Method

*Data* . Data were collected in 1995, during the entire second semester (January through May). Within each class in the evaluation project, portfolios from 6 students were collected according to a stratified (gifted, nonlabeled, and educational support) random sample. There were portfolio data from a total of 313 students, grades 1-10, in whose portfolios were included a total of 2862 annotated work samples. Portfolios were from mathematics (88%) and language arts (12%) classes. Fifty-two percent of students were male; 48% were female. Their status was 27% gifted, 47% nonlabeled, and 26% educational support. Data from student annotation forms and teacher ratings of achievement and time use were the same as for Study One, except that the SE variable (DIFF times DOMORE) was not calculated. The SE variable was dropped because of the results of the validity analyses in Study One.

***Analyses.*** Two reliability analyses were done, to replicate those conducted for Study One. (1) Cronbach's alpha was calculated for DIFF and DOMORE ratings, treating annotations within one student's portfolio as items on the respective scales. Student was the unit of analysis. The rationale for this analysis was the same as for Study One, namely, that overall DIFF and DOMORE ratings for each student had been used in the evaluation and thus internal consistency was required. (2) Generalizability studies with a Person by Time design were conducted for DIFF and DOMORE scales. The rationale was again the same as for Study One, namely, to establish how many annotations were required for reliable overall scores and to determine whether repeated use of the same form compromised the validity of the information.

The external structure of DIFF and DOMORE ratings were again examined as validity evidence. A replication of the qualitative study of validity evidence, coding student written responses, was not possible within the time frame of the study.

## Results and Discussion

***Reliability.*** Internal consistency results for DIFF were slightly higher than those found in Study One. Alpha reliability ranged from.55 for 6 annotations per portfolio to.70 for 9 annotations (see Table 2.1,cf. .59 to .64 for Study One). Internal consistency results for DOMORE were slightly lower than those found in Study One. Reliability ranged from .65 for 7 annotations per portolio to .72 for 10 annotations (see Table 2.1, cf. .66 to .80 for Study One ). The number of students upon which these analyses were based was larger in Study Two than in Study One, and the estimates are therefore more stable.

**Table 2.1** Reliability (Internal Consistency) of Scales for Six through Ten Annotations per Portfolio

| Scale | k | α | n |
|---|---|---|---|
| DIFF | 6 | .55 | 274 |
| | 7 | .63 | 240 |
| | 8 | .67 | 207 |
| | 9 | .70 | 185 |
| | 10 | .66 | 157 |
| DOMORE | 6 | .66 | 248 |
| | 7 | .65 | 217 |
| | 8 | .70 | 194 |
| | 9 | .70 | 171 |
| | 10 | .72 | 144 |

The generalizability results from Study Two were virtually identical to those from Study One (see Tables 2.2.and 2.3). The conclusion that repeated use of the same form over time did not have any effect was once again supported. Less than 1% of the variance in both DIFF and DOMORE was due to Time (Table 2.2). The sizes of the other variance components was similar to those in Study One. The finding that the Person-by-Time interaction was the largest effect for both DIFF and DOMORE was replicated. As in Study One, this large interaction term is more similar to results from performance assessments than from paper-and-pencil tests.

Both DIFF and DOMORE scales generalized at an acceptable level over repeated use in a portfolio (Table 2.3). As for Study One, relative and absolute generalizability coefficients were similar because the Time effect was almost zero. Generalizability coefficients for DIFF were virtually identical to those from Study One; generalizability coefficients for DOMORE were slightly lower than those for Study One (see Table 1.4).

**Table 2.2** Variance Component Estimates for Scales, for Person x Time Design

| | DIFF | | DOMORE | |
|---|---|---|---|---|
| Source of Variability | Estimated Variance Component | Percent Total Variability | Estimated Variance Component | Percent Total Variability |
| Person | .0955 | 19% | .0416 | 21% |
| Time | .0017 | < 1% | .0003 | < 1% |
| P x T | .3987 | 80% | .1573 | 80% |

DIFF variance estimates based on 240 persons, 7 times
DOMORE variance estimates based on 216 persons, 7 times

**Table 2.3** Generalizability of Scales, for Six through Ten Annotations per Portfolio

| Scale | k | $\hat{\rho}^2$ | $\hat{\phi}$ |
|---|---|---|---|
| DIFF | 6 | .59 | .59 |
| | 7 | .63 | .63 |
| | 8 | .66 | .66 |
| | 9 | .68 | .68 |
| | 10 | .71 | .70 |
| DOMORE | 6 | .61 | .61 |
| | 7 | .65 | .65 |
| | 8 | .68 | .68 |
| | 9 | .70 | .70 |
| | 10 | .73 | .72 |

DIFF estimates based on 240 persons, 7 times
DOMORE estimates based on 216 persons, 7 times

***Validity*** Table 2.4 shows that the correlational validity evidence from Study Two was in the same direction as for Study One but was stronger. Both DIFF and DOMORE were significantly but weakly related to the percent of instructional objectives attained (of those attempted for the portfolio, PCTACH) and to teacher' judgments of students time use (TIME) and evaluation of achievement quality (EVAL). Teachers judge more efficient time use for students who rated their work samples less difficult overall. Teacher judged more efficient time use for students who indicated more interest in doing their work. These findings accord with the significant results from Study One, but offer slightly stronger evidence that students' in-class behavior, use of lesson and study time, is consistent with their annotations. For Study Two, unlike Study One, there were also significant but weak relationships between the quality of students' work (percent achievement and evaluation) and the DIFF and DOMORE variables. More difficult ratings on annotations were associated with lower percentages of achievement of attempted objectives and with lower teacher evaluations of work quality. Willingness to do more of the same kind of work was associated with higher percentages of achievement of attempted objectives and with higher teacher evaluations of work quality (Table 2.4). Study Two's results may be more accurate because of the length of data collection . In 1995, the school district required that portfolios be kept for the entire second semester. In1994 ( Study One's data), the length of collection varied from a few weeks to an entire semester, depending on the class.

**Table 2.4** Correlations between Scales and Selected External Measures

| | PCTACH | TIME | EVAL |
|---|---|---|---|
| DIFF | -.15 * | -.26 ** | -.13 * |
| DOMORE | -.14 * | -.18 * * | .12 * |

* $p < .05$, **$p < .01$, n ranged from 242 to 284 students
PCTACH = percent of learning objectives attempted within the portfolio that were achieved
TIME = efficiency of time use, 0 = inefficient, 1 = efficient, average over all learning objectives represented in the portfolio
EVAL = teacher evaluation of achievement, on a scale of 1 to 4, averaged over all learning objectives represented in the portfolio

In Study One, lack of relationship between annotations and achievement was interpreted to mean that the annotations measured something different from achievement. This intepretation may still be the one that makes the most sense. Relationships between annotations and achievement were significant (and logical, since the annotations were affixed to school work samples meant to demonstrate achievement); nevertheless, they were small enough that they do not explain much of the variance in achievement. Clearly, additional information, besides achievement, is contained in the DIFF and DOMORE ratings.

## Conclusion

The basic indicators on the Student Annotation Form, the DIFF and DOMORE questions, did stand up to reliability and validity study. Reliability and validity of the items on the Student Annotation Form were confirmed for the purpose of a districtwide evaluation, where efficient and aggregatable measures were required. An attempt to construct a multiplicative self-efficacy measure from these ratings did not stand up to a validity study. The Student Annotation Form provided a simple, useful tool for indicating student judgments about individual work samples in a portfolio. Aggregated over time, during the longitudinal process of keeping a portfolio, these annotations also provided simple but useful measures of students' critical judgments of their work output--their perceptions of its difficulty and their willingness to do more of it. These are outcomes of general interest in current educational practice.

## References

Azar, M., Cass, M. Cipieliewski, J., Jordan, R., Kraetke, S., & Markel, K. (1993). ***Sit beside me...Macomb portfolio project. 1990-93.*** Clinton Township, MI: Macomb Intermediate School District. (ERIC Document Reproduction Service No. ED 361 385)

Brennan, R. L., Gao, X., & Colton, D.A. (1995). Generalizability analyses of Work Keys listening and writing tests. ***Educational and Psychological Measurement, 55,*** 157-176.

Calfee, R. C., & Perfumo, P. (1993). ***Student portfolios and teacher logs: Blueprint for a revolution in assessment.*** Berkeley, CA, & Pittsburgh, PA: Center for the Study of Writing. (ERIC Document Reproduction Service No. ED 362 887)

Farnan, N., & Kelly, P. (1991). Keeping track: Creating assessment portfolios in reading and writing. ***Reading, Writing, and Learning Disabilities, 7,*** 255-269.

Ferguson, S., & Maples, C. (1992). Portfolios: Windows on learning: Zeroing in on math abilities. ***Learning, 21*** (3), 38-41.

Goodman, K.S., Bird, L. B., & Goodman, Y.M. (1992). ***The whole language catalog supplement on authentic assessment.*** Santa Rosa, CA: American School Publishers.

Herman, J.L., & Winters, L. (1994). Portfolio research: A slim collection. ***Educational Leadership,*** 52 (2), 4 8 - 55.

Klein, S.P., McCaffrey, D., Stecher, B., & Koretz, D. (1995). The reliability of mathematics portfolio scores: Lessons from the Vermont experience. ***Applied Measurement i n Education, 8,*** 243 - 260.

Roe, M., & Vukelich, C. (1995,April) .***That was then and this is now: A longitudinal study of teachers's portfolio practices.*** Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Salomon, G. (1984). Television is "easy" and print is "tough": The differential investment of mental effort as a function of perceptions and attributions. ***Journal of Educational Psychology, 76.*** 647 - 658.

Wang, M. C. (1992). ***Adaptive education strategies: Building on diversity. Baltimore:*** Paul H. Brooks Publishing.

Zimmerman, B. J., & Martinez-Pons, M. (1992) Perceptions of efficacy and strategy use in the self-regulation of learning. In D.H Schunk & J.L. Meece (Eds.), ***Student perceptions in the classroom*** (pp. 185 - 207). Hillsdale, NJ: Lawrence Erlbaum.