DOCUMENT RESUME

ED 390 921 TM 024 331

AUTHOR Cope, Ronald T.
TITLE Cautionary Observations on Reliability and Equating of Forms in High Stakes Performance Assessment: The Problem of Granularity.
PUB DATE Apr 95
NOTE 25p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, CA, April 19-21, 1995).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Difficulty Level; Educational Assessment; *Equated Scores; Performance Based Assessment; *Sample Size; Scaling; *Statistical Analysis; Test Construction; Testing Programs; *Test Reliability; *Writing Evaluation
IDENTIFIERS *Granularity (Statistics); *High Stakes Tests

ABSTRACT

This paper deals with the problems that arise in performance assessment from the granularity that results from having a small number of tasks or prompts and raters of responses to these tasks or prompts. Two problems are discussed in detail: (1) achieving a satisfactory degree of reliability; and (2) equating or adjusting for differences of difficulty among tasks or prompts. Empirical results from the Schul and Linacre (1995) study of writing assessment are used to amplify the discussion of reliability problems, and a set of hypothetical equating results are used to illustrate the problem of trying to make adjustments for difficulty when there are few scale points. The discussion concludes with suggestions for how performance assessment programs might attempt to deal with the problems that arise from high degrees of granularity. (An appendix contains a derivation of statistics used in the figures.) (Author)

Cautionary Observations on Reliability and Equating of Forms in High Stakes Performance Assessment: The Problem of Granularity

Ronald T. Cope

American College Testing

A paper presented at the annual meeting of the National Council on Measurement in Education

San Francisco

April, 1995

Abstract

This paper deals with the problems that arise in performance assessment from the granularity that results from having a small number of tasks or prompts and raters of responses to these tasks or prompts. Two problems are discussed in detail: (1) achieving a satisfactory degree of reliability, and (2) equating or adjusting for differences of difficulty among tasks or prompts. Empirical results from the Schulz and Linacre (1995) study of writing assessment are used to amplify the discussion of reliability problems, and a set of hypothetical equating results are used to illustrate the problem of trying to make adjustments for difficulty when there are few scale points. The discussion concludes with suggestions for how performance assessment programs might attempt to deal with the problems that arise from high degrees of granularity.

Performance assessments such as direct writing assessments are typically granular. Each examinee receives one to three tasks or prompts. Two or three raters grade the responses on a scale such as 1-4 or 1-5. The resulting scale of total raw scores will range from possible scores of 2-8 when there is 1 prompt, 2 raters each scoring the response 1-4, up to scores of 9-45 (3 prompts, three raters, responses rated 1-5). The resulting raw score scales have the same number of distinct points as do dichotomously scored multiple-choice tests having from 6 to 36 items. In these examples, granularity arises from two sources: (1) a small number of tasks, stimuli, or prompts, and (2) few raters. Although granularity is to some degree present in all educational measures that produce integer scores, it is more obvious and much more of a problem in the situations discussed in this symposium.

A pronounced degree of granularity causes problems in the following areas:

1. Attaining satisfactory reliability

2. Attempting to adjust scores on different forms (single prompts or sets of prompts) for variations in difficulty

Let us consider these points in detail.

1. Achieving satisfactory reliability in the face of granularity. Ten or more years ago, posing the question why is it hard to attain a satisfactory degree of reliability when the responses to one or two writing prompts are graded by one or two raters? would likely have evoked the answer, "Obviously because there are too few raters. Research has shown that teachers' grading of essays and essay examinations is highly inconsistent, even when the same teacher grades a set of papers on different occasions. The only way to attain a satisfactory degree of reliability is to use as many trained raters as possible."

More recent research, however, contradicts this wisdom of yesteryear (see, e.g., Linn and Burton, 1994). It turns out that two or three properly trained and "calibrated" raters are usually sufficient. Additional raters are a luxury; they likely will not greatly improve reliability.

The greater source of unreliability and measurement error turns out to be having too few items, prompts, or individually scorable units. Why should this be such a strong source of measurement error? Certain possibilities present themselves: one possibility is that the small number of prompts or items greatly limits the coverage of the domain of knowledge or skill. Also, the process of responding to the item or prompt may be poorly sampled. Consider the student who responds to a single writing prompt. Despite a generally high level of writing skill, the student may lack knowledge of the subject of the prompt. Often an attempt is made to set prompts that are very general in nature in the hope that all examinees can respond knowledgeably, but the hope that such

prompts are easy to create may be unrealistic. We could also hope that a good writer's skill will show even through an uninspired response to a particular prompt, but this may not be the case.

Figures 1, 2, and 3, which use statistics from the Schulz and Linacre study (1995), illustrate the comparative effects of increasing the number of writing prompts versus increasing the number of raters (See Appendix for derivation of statistics used in the figures). In Figure 1, we see that using two raters instead of one results in a sizable increase in reliability, from the mid- .70's to the mid- .80's. But increasing the number of raters from two to three results in a much smaller increase in reliability, and using more than three raters results in only very slight further increases.

In contrast, Figure 2 shows sizable increases of reliability from adding a second and third prompt, and this advantage holds as the number of ratings and raters increases. Figure 3 shows between-prompt reliability as a function of the ratio of student variance to the variance from the interaction of student and prompt. Again, adding a second and third prompt results in a sizable increase of reliability even when student variance greatly exceeds the student-by-prompt interaction.

2. Problems of trying to adjust for difficulty. A pronounced degree of granularity also makes the traditional adjustments for form difficulty problematic. I shall consider two aspects of the problem: (1) regarding examinee true scores as equal under equating transformations, (2) trying to adjust scores when there are few points and big "quantum jumps" in the score scale.

We first need to consider an important aspect of the process of equating alternate test forms. Equating of multiple-choice tests is a group-referenced process: it does not tailor adjustments for form difficulty to individual examinees. A single conversion function from one form to another is applied to the scores of all examinees who take the form whose scores are to be converted. In the traditional setting, alternate forms of tests typically consisting of at least 20 items are regarded as having perfectly correlated true scores; that is, the various forms are assumed to measure the identical trait or combination of traits. This is a reasonable assumption if the forms are constructed under a single set of content and process specifications. If this condition is satisfied, we can suppose that a single examinee's true score on each form is the same under raw score scale transformations that allow for variations of difficulty. The basic supposition here is that under the equating transformation it should make no difference which form any examinee takes.

But now consider what happens with forms that consist of only one or two items or prompts. In this case the assumption of equal true scores is questionable: considerable prompt-examinee interaction is to be expected because a given examinee will likely know more about the subject matter of or be more interested in or otherwise more able to respond to one prompt than another; that is, the content domain will be poorly sampled. Thus the assumption of equal true scores is hard to defend: even if we hold to the equal true scores assumption, the observed score is subject to a large error of measurement arising from the examinee-form interaction. In the traditional equating situation involving tests with a much larger

number of scorable units or items, we can expect these examinee-prompt interactions to cancel out; with an increased of number c items, the error variance arising from examinee-prompt interactions becomes a smaller and smaller proportion of total variance. As a result, the equating conversion should apply satisfactorily to any and all examinees. Conversely, with a decrease of number of items, the error variance from examinee-prompt interactions becomes an increasingly greater proportion of total variance. And such a proportional increase of error implies decreasing reliability. This brings us right back to the problem of low reliability when there are few scorable units.

Now consider what happens when the score scale consists of only a few integer points. Table 1 illustrates what might happen in trying to equate four forms with raw score scales ranging from 2 through 8. For converting scores of Form A and Form B, we are in luck. Except for the top and bottom scores, each Form A score has an exact cumulative percent counterpart in Form B. It is reasonable to take a 4 on Form A to be equivalent to a 3 on Form B, a 5 on Form A to be equivalent to a 4 on Form B, and so on. But of course we are unlikely to obtain cumulative percents that correspond so neatly. Instead, we might get the relationship of Form A and Form C. For a given raw score, Form C has a slightly higher percent of examinees at that score or lower. Thus throughout its score range, Form C is slightly harder than Form A. Accordingly, we would like to make slight downward adjustments to Form A scores or slight upward adjustments to Form C scores. Unfortunately, the limited number of integer score points does not permit such adjustments. The situation

is even worse with Form A and Form D. Cumulative percents for Form D suggest that Form A scores would convert to Form D scores about halfway between integer values. A score of 5 on Form A might convert to a 4.5 on Form D. Now what do we do? It makes a big difference whether a 5 on Form A is converted to a 4 or a 5 on Form D--especially if we are trying to set equivalent cut scores on these forms. Granularity has again created a serious problem.

What about converting raw scores to a scale with more points, say 0-100? This will not solve the problem, only disguise it. For any one of the four forms, most of the scores on the new scale will be unattainable. Abandoning raw scores for a logit scale will only replace the widely-separated integer scores with separate "probability-frequency bumps."

The problems created by high granularity are of greatest concern in high-stakes situations. How to remedy the problems? Here are some approaches to consider.

1. Increase the reliability and number of raw score scale points by having more prompts/forms and raters, with priority given to increasing the number of prompts to achieve greater content domain coverage. Strongly prefer the use of two prompts over one. Practical limitations will of course prevent arbitrarily large increases in the number of prompts.

2. Consider the use of supplemental measures, such as portfolios, traditional standardized measures, and class grades or other teacher-generated assessments of writing competence. Note that at least one state, California, has mandated use of multiple measures for placement testing at the college level (J. Roth, personal communication, March 15, 1995). Care must be taken to ascertain whether additional measures are reliable enough to increase instead of decrease overall reliability and whether they are dimensionally close enough to the base measure not to alter seriously the nature of what is assessed.

3. Consider increasing the number of rating scale points for each scorable unit. This should alleviate the problem of trying to equate scales of few score points.

4. Allow unsuccessful examinees to retest on a different form after a suitable time interval.

5. Consider restricting the use of performance writing assessments to evaluation of students by teachers and to group (class, school, district) rather than individual assessments.

References

Linn, R.L., & Burton, E. (1994). Performance-based assessment: Implications of Task Specificity. *Educational Measurement: Issues and Practice,* **13**(1), 5-8, 15.

Schulz, E.M., & Linacre, J.M.(1995). *A comparison of the many-facet Rasch model and generalizability procedures for person measurement in writing assessment.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Table 1.

Cumulative Percentages of Raw Scores on Four Forms of a Hypothetical Writing Assessment

	Cumulative Percent of Scores			
Raw Score	Form A	Form B	Form C	Form D
8	100	100	100	100
7	94	98	95	97
6	80	94	82	87
5	53	80	57	66
4	31	53	33	42
3	7	31	8	19
2	2	7	3	4

Note. Range of possible raw scores is 2-8. Large equivalent groups of examinees are assumed to have taken the forms.

Figure 1: Within-Prompt Reliability vs. Number of Raters

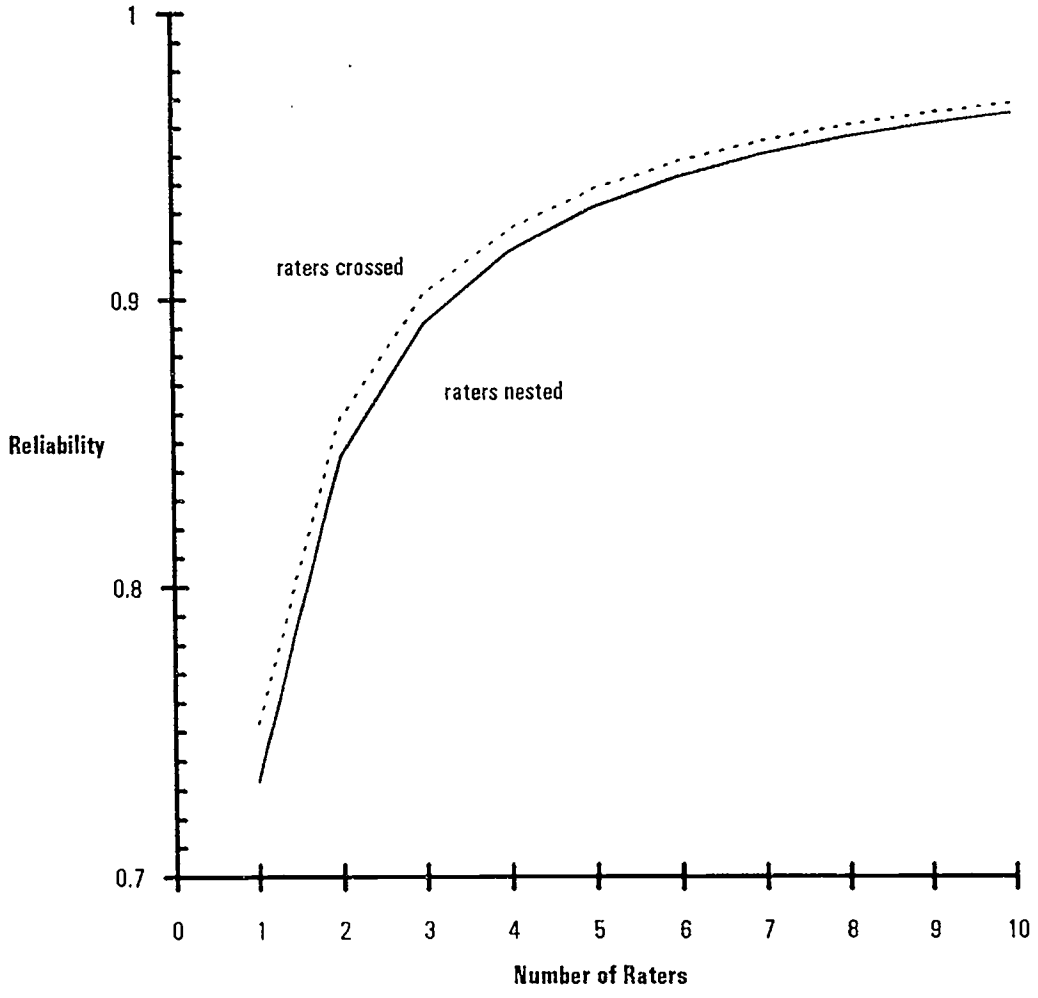


Figure 2: Between-Prompt Reliability

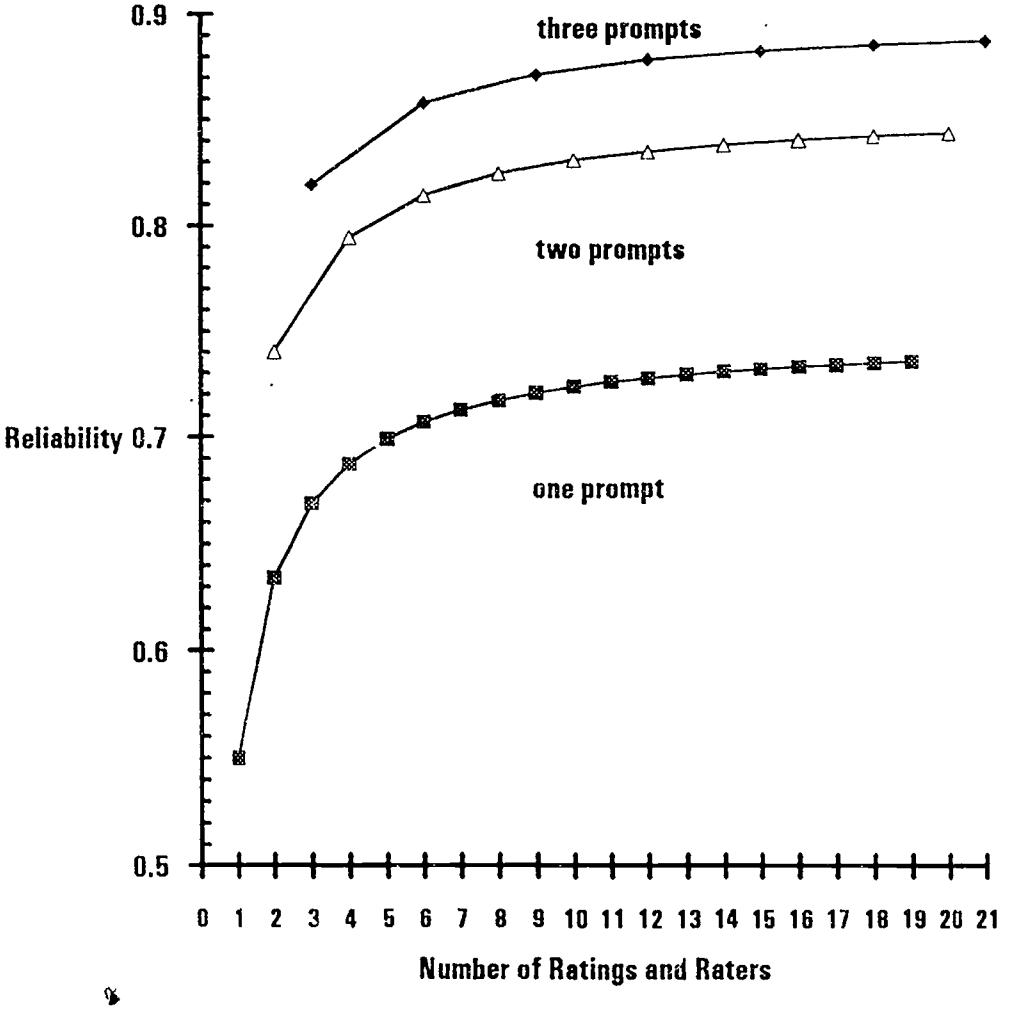
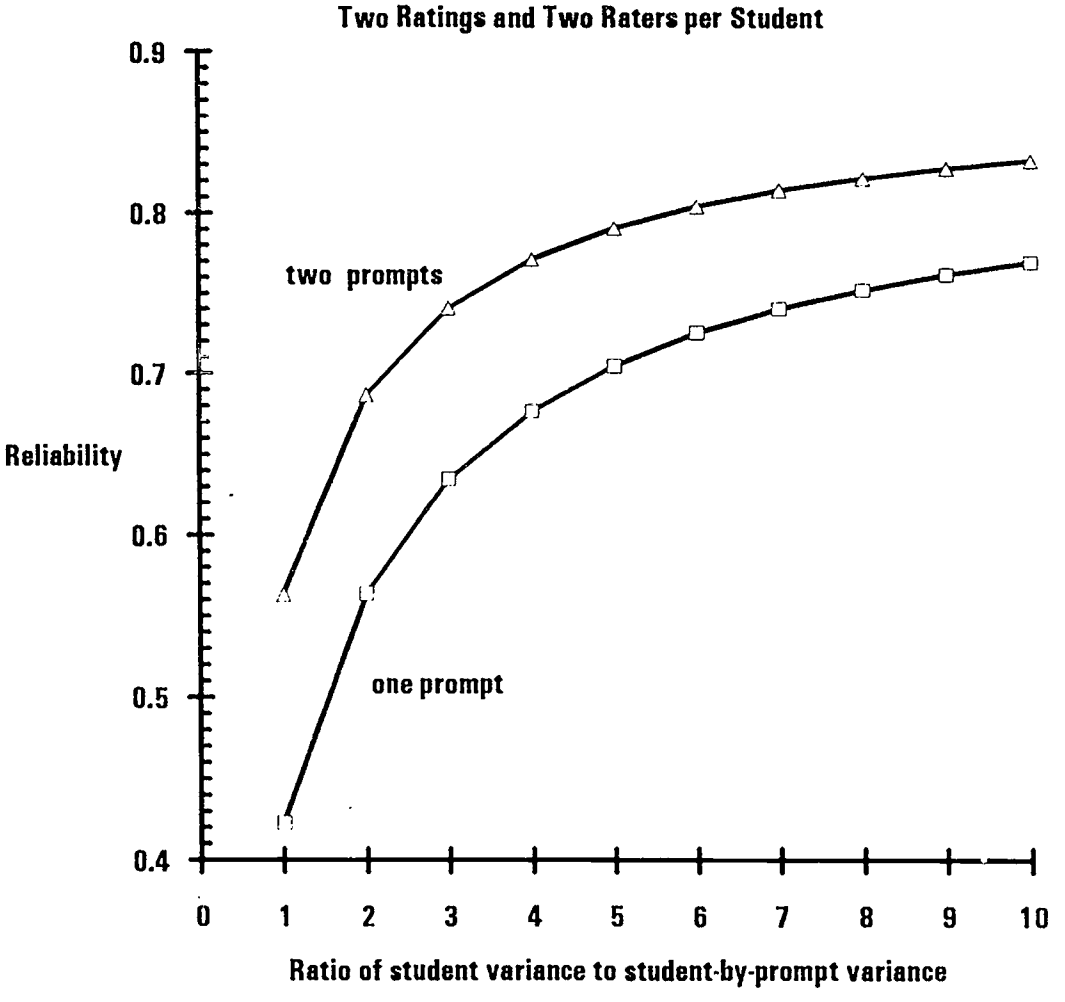


Figure 3: Between-Prompt Reliability



Appendix

Derivation of Statistics

Used in Figures 1,2, and 3

To illustrate the effects of subject-by-prompt interactions on the reliability of writing assessment scores, the reliability of differences among students within and across prompts was estimated. For all estimates, it was assumed that a student supplies only one writing sample per prompt. In a writing assessment consisting of students, prompts and raters, there are seven separable components of variance:

σ_s^2 variance due to differences between students,

σ_r^2 variance due to differences between raters,

σ_p^2 variance due to differences between prompts,

σ_{sr}^2 variance due to the interaction of students with raters,

σ_{sp}^2 variance due to the interaction of students with prompts,

σ_{rp}^2 variance due to the interaction of raters with prompts,

$\sigma_{e,srp}^2$ the combination of error variance and student-by-rater-by-prompt interaction effects,

The general form for the reliability of difference between two students is:

$$R = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_E^2} \quad (1)$$

where σ_T^2 is true variance and σ_E^2 is error variance. The definition of σ_T^2 and σ_E^2 depend on whether the students to be compared have prompts and raters in

common.

For the reliability of differences between students who have taken a single prompt, but to whom one or more rater(s) were assigned at random from a large pool of raters (raters nested within students), σ_T^2 is:

$$\sigma_T^2 = \sigma_{s,sp,srp}^2 = \sigma_s^2 + \sigma_{sp}^2 + \sigma_{srp}^2(\frac{1}{n_r}) \quad (2)$$

and σ_E^2 is:

$$\sigma_E^2 = \frac{\sigma_r^2 + \sigma_{rp}^2 + \sigma_{e,sr}^2}{n_r} \quad (3)$$

This definition of error variance is a close, but conservative approximation for reliability of differences between students in a large scale assessment who might, but are not likely to have, one or more raters in common.

For the reliability of differences between students who have taken a single prompt and share the same rater(s) (raters crossed with students), the definition of σ_T^2 is as given in equation 2, but the error term is:

$$\sigma_E^2 = \frac{\sigma_{e,sr}^2}{n_r} \quad (4)$$

For the reliability of differences between students to whom both prompts and raters are assigned at random from large pools of raters and prompts, σ_T^2 is:

$$\sigma_T^2 = \sigma_s^2 \tag{5}$$

and σ_E^2 is:

$$\sigma_E^2 = \frac{\sigma_r^2}{n_r} + \frac{\sigma_p^2}{n_p} + \frac{\sigma_{sr}^2}{n_r} + \frac{\sigma_{sp}^2}{n_p} + \frac{\sigma_{rp}^2}{n_r n_p} + \frac{\sigma_{e,srp}^2}{n_r n_p} \tag{6}$$

Equation 6 is equivalent to the definition of the *absolute* error term for an assessment in which students prompts and raters are fully crossed (Shavelson and Webb, 1991, p 89). Here it is suggested as a reasonable definition of the *relative* error term when raters are nested within prompts and prompts nested within students. This hypothetical, "nested universe" applies approximately to two randomly selected students from the data accumulated in a large scale annual writing assessment in which different prompt(s) are used each year, and raters are randomly assigned to students from a large and fluid rater pool. The odds are small that any two such randomly selected students would have a prompt or rater in common.

The variance estimates reported by Schulz and Linacre (this symposium) were obtained from an r*(s:p) design (raters crossed with subjects nested within prompts) and were:

$\hat{\sigma}^2_{s,sr,srp} = .546$

$\hat{\sigma}^2_r = .011$

$\sigma^2_r = .009$, and

$\hat{\sigma}^2_{e,sr} = .179$

$\hat{\sigma}^2_{s,sr,srp}$ was defined in equation 2, with $n_r=2$.

Estimates from the Schulz and Linacre study were used to compute reliability of differences among students *within* prompts (equation 2, for σ^2_T) when raters are either nested within students (equation 3 for σ^2_E) or crossed with students (equation 4 for σ^2_E).

Figure 1 shows within-prompt reliability (subjects nested within one prompt, $n_p=1$) as a function of the number of raters ($n_r=1, 2, ..., 10$) when raters are nested within students ("different rater(s) for each student") or crossed with students ("same rater(s) for all students"). Given the estimates of σ^2_r and σ^2_{rp} in Schulz and Linacre, within-prompt reliability is only slightly affected by whether all students are rated by the same or different raters. The difference in the reliability of within-prompt differences between students would be greatest if there were only one rater/rating per student (.75 for raters crossed versus .73 for raters nested).

For estimating the reliability of differences across prompts, separate estimates of each of the variances in equations 5 and 6 are needed. The only design that can provide

all seven of these estimates is one in which subjects, prompts and raters are crossed. In order to illustrate the possible effects of subject-by-prompt interactions, however, a number of assumptions were used to allocate the values reported in Schulz and Linacre to the seven variance components in equations 5 and 6.

The estimate, $\hat{\sigma}^2_{s,sr,srp}$=.546 (Schulz and Linacre) was attributed exclusively to $\hat{\sigma}^2_s + \hat{\sigma}^2_{sp}$. [The magnitude of $\sigma^2_{srp}/2$ (equation 2, given n_r=2 in Schulz and Linacre) was assumed to be relatively small and inconsequential for demonstrating the impact of subject-by-prompt effects on reliability.] The ratio,

$$X = \frac{\sigma^2_s}{\sigma^2_{sp'}} \tag{2}$$

was identified as the key quantity for illustrating the impact of σ^2_{sp} on the reliability of differences across prompts. Specific values of X, (e.g., X=3) were assumed for computing reliability coefficients. Given the estimate, $\hat{\sigma}^2_s + \hat{\sigma}^2_{sp'}$=.546:

$$\hat{\sigma}^2_s = .546 - \frac{.546}{X+1} \tag{3}$$

and

$$\hat{\sigma}^2_{sp'} = \frac{.546}{X+1} \tag{4}$$

Another assumption was that $\sigma^2_{p'} = 0$. This assumption is equivalent to assuming that

prompts that have been perfectly equated on a group basis, and it is the reliability of equated scores that is being estimated. Note that perfect group-based equating does not imply that there is no student-by-prompt interaction ($\hat{\sigma}^2_{sp}>0$).

Finally, the variance estimate $\hat{\sigma}^2_{e,sr}$=.179 (Schulz and Linacre) was divided evenly between $\hat{\sigma}^2_{sr}$ and $\hat{\sigma}^2_{e,srp}$, so that $\hat{\sigma}^2_{sr} = \hat{\sigma}^2_{e,srp} = .09$. The assumptions underlying these estimates were 1) $\sigma^2_{srp}\approx 0$, and 2) $\hat{\sigma}^2_{sr} = \hat{\sigma}^2_{e}$. The effect of the latter assumption on the apparent effect of subject-by-prompt interactions (which was of primary interest in this study) was ascertained by computing reliabilities corresponding to the extreme cases, 1) $\hat{\sigma}^2_{sr}$=0, $\hat{\sigma}^2_{e,srp}$=.179, and 2) $\hat{\sigma}^2_{sr}$=.179, $\hat{\sigma}^2_{e,srp}$=0.

Table A1 summarizes the variance estimates used to compute reliability of subjects across prompts.

Table A1

Source of Variance	Notation	Estimate
Subjects	σ^2_s	.546-(.546/(X+1))
Raters	σ^2_r	.011
Prompts	σ^2_p	0
SxR	σ^2_{sr}	.09
SxP	σ^2_{sp}	.546/(X+1)
RxP	σ^2_{rp}	.009
error,SxRxP	$\sigma^2_{e,srp}$	.09

Figure 2 shows reliability as a function of the number of ratings=raters per student (n_r) under three conditions: 1, 2 or 3 prompts nested within students. The ratio, $X=\sigma^2_s / \sigma^2_{sp}$ (see Table A1) was assumed to be 3. [Reliability estimates for 1 nested prompt per student did not depend on how the variance estimate, $\hat{\sigma}^2_{e,sr}$=.179 from Schulz and Linacre was divided up because the divisors for $\hat{\sigma}^2_{sr}$ and $\hat{\sigma}^2_{e,srp}$ in equation 6 are equal when n_p=1.]

Comparing specific plots in Figures 1 and 2 shows how much greater reliability is within prompts than between prompts. When X=3 and there is one prompt per student, within-prompt reliability starts at .73 with 1 rater (nested within students)

and approaches 1 as the number of raters increases; between-prompt reliability starts at .55 with 1 rater and approaches a maximum of .75 as the number of raters (=ratings) increases.

The plots in Figure 2 show how much the reliability of differences between students over prompts can be increased by increasing the number of prompts per student, with the number of both raters and ratings held constant. For example, the reliability of three ratings from three raters rating three prompts (.82) is much greater than the reliability of three ratings from three raters rating a single prompt (.67). The total number of ratings and raters is the same in each case (three). If only the number of ratings, but not raters were held constant, results could differ considerably. For example, one rater could provide all three ratings of three prompts presented to a student. In such a case, a rater-by-student interaction effect could offset the gain in reliability expected from increasing the number of prompts. More precise information is needed concerning the relative magnitude of two way interaction effects in the assessment.

Figure 3 shows between-prompt reliability as a function of the ratio, $X=\sigma_s^2 / \sigma_{sp}^2$, when there are two raters and two ratings per student, but one or two prompts. Prompts and raters are again assumed to be nested within students. When there is one prompt, the two ratings apply to the same prompt. When there are two prompts, there is only one rating/rater per prompt. Compared to just one prompt, between

prompt reliability is considerably higher for two prompts, even for high values of $\sigma_s^2 / \sigma_{sp}^2$.

To see what the effect of assuming $\hat{\sigma}_{sr}^2 = \hat{\sigma}_{e,srp}^2 = .09$ (see above) had been on the trends for multiple prompts in Figures 2 and 3, these plots were generated again using the alternative extreme assumptions that 1) $\hat{\sigma}_{sr}^2=0$, $\hat{\sigma}_{e,srp}^2=.179$, and 2) $\hat{\sigma}_{sr}^2=.179$, $\hat{\sigma}_{e,srp}^2=0$. The trends in the plots were not substantially altered.