

Using the SAS System to Assess the Reliability of Gain Scores

Jim Penny, University Research Associates, Inc., Jamestown, NC

ABSTRACT

The effect of pre-test reliability, post-test reliability, pre- and post-test correlation, and changes in pre- and post-test variance on the reliability of gain scores is presented graphically, with gain score reliability computed using empirical data derived from a series of Monte-Carlo studies based on the IRT model. Conditions where the gain score reliability may be sufficiently high for use in assessment, though difficult to achieve, are identified and discussed, as are converse conditions, which are perhaps more common, where gain score reliability may be sufficiently low enough to jeopardize inference.

INTRODUCTION

After total score, there is may be no simpler, or more natural, composite used in educational measurement than the gain score. Ask nearly educator what an appropriate index of increased learning might be for a given situation, and the response will quite likely include the improvement of a post-test score over a pre-test score. Ask that same educator what she might estimate the reliability of that gain score to be, and the response will most likely be something on the order of the reliabilities of the pre- and post-test, perhaps an average of the two. That response is very likely wrong.

The reliability of gain scores, a subject that is singularly interesting from a historical perspective, has been the focus of some study and no little debate for more than 4 decades, beginning with Gulliksen (1950), continuing with Lord & Novick (1968), Cronbach & Furby (1970), O'Connor (1972), Overall & Woodard (1975), Lim & Slinde (1977), Zimmerman & Williams (1982) and Rogosa & Willett (1983), and more recently with Willett (1988), Humphreys (1991, 1993), Zimmerman, Williams, & Zumbo (1993), Zimmerman (1994), and Feldt (1995). Using Classical True Score theory (Crocker & Algina, 1986), each of these studies explores in part the effect on gain score reliability of pre-test reliability, post-test reliability, changes in reliability from pre-test to post-test, pre-test and post-test correlation, changes in variance from pre-test to post-test, and, finally, the use of true score estimates.

The combined results of these studies have shown that the reliability of gain scores can be manipulated substantially, resulting in situations where the difference score is sufficiently reliable for the purposes of educational assessment, in situations where the reliability is sufficiently low as to compromise a study, and in middle-ground situations where the reliability is questionable but perhaps sufficient. The factors that tend to increase the reliability of gain scores are (1) increasing the reliability of the pre- and post-tests, (2) increasing the reliability of the post-test over that of the pre-test, (3) increasing the variance of the post-test over that of the pre-test, and (4) reducing the correlation between the pre- and post-test. As a result, the reliability of gain scores has, across four decades, become well-understood, and frequently derided, by psychometric specialists, though under what conditions the use of these scores may compromise the interpretation of a study will most likely continue to prompt substantial discussion.

Measurement practitioners, however, may not be as disposed to follow these discussions, the algebraic manipulations, and the tables of estimated reliabilities; their interests, perhaps motivation, may lie elsewhere. Consider for instance, the chair who must persuade her administration that an expensive new writing curriculum differentially improves writing scores beyond those of the less-expensive traditional program, or the developer of an SAT preparation program that purports to boost scores on the SAT; both are quite likely to consider an analysis of gain scores. Indeed, there is perhaps no measurement more "natural" than the gain score in situations such as these; hence, it is essential for those who would assess a program using gain scores to know the conditions under which this measure of improvement is reliable, and while it is certainly true that other issues such as differential performance, restriction of range, memorization, and subject mortality probably present the

measurement professional with problems greater than those presented by gain score reliability, it is without question that if change is to be assessed and the gain score has been chosen as a component of the assessment then factors that control the behavior of gain scores and their effect of gain score reliability must be understood well.

The web of relationships responsible for the reliability of gain scores, while certainly not exhibiting the complexity of some statistical frameworks such as Item Response Theory or Generalizability Theory, can certainly present a daunting learning curve to the practitioner who is in the process of planning a study where the analysis of gain scores would seem, at least on the surface, a reasonable undertaking. However, to expect that practitioner, perhaps puzzled by all the ostensibly contradictory statistical advice, to sort through the findings of a dozen or more research papers strewn through 45 years of measurement journals and texts is perhaps asking a bit much, especially when the subject seems as innocuous, indeed safe, as the simple difference of pre- and post-test scores.

To bring into focus these relationships, to examine all the technical factors that influence difference score reliability, and to examine the congruency of the analytical definitions of gain score reliability with data derived under tightly controlled circumstances and presented within a single arena was the purpose of this research; and in doing so, we will be able to present the measurement professional with the tools necessary to determine whether or not the difference score can be a meaningful component of an educational assessment.

METHODOLOGY

Item Response Theory (IRT) was used to provide the framework for a series of Monte-Carlo studies of classical reliability; however, this is not an IRT study. IRT only provides the context with which we model classical test theory to study the reliability of gain scores. Indeed, IRT provides no notion of reliability within the classical sense of the word, offering instead a construct of test "information." Beyond taking advantage of the high degree of correlation between total score, as defined by number of items correct, and the latent IRT ability, which we use as an implicit proxy for the classical "true" score, these Monte-Carlo studies are purely classical in nature, affording the opportunity to manipulate the factors that influence gain score reliability across ranges of values that would be difficult to obtain reliably with "real" data.

The basis of this study was a 100 item Monte-Carlo test created using the 3-parameter Item Response Theory model. The values of the discrimination parameter were selected from a stream of pseudo-random numbers, $\sim N(1, 0.3)$. The values of the difficulty parameter were selected from a similar stream, $\sim N(0, 1)$. The values of the pseudo-guessing parameter were selected from a third stream, $\sim N(0.20, 0.05)$. The probability that a subject would correctly answer an item was computed using the item parameters and the subject ability, the latter to be described in detail in the next section. To mark the item correct or incorrect, a random number was drawn from a flat distribution bounded by 0 and 1. If the probability of a correct response exceeded the number chosen from the flat distribution, the item was coded as "correct;" otherwise, it was coded as "incorrect." The total score of a subject on the test was the number of items coded as correct.

Four streams of random numbers, $\sim N(0, 1)$, were drawn and manipulated to provide estimates of subject IRT ability in four subject groups. The first group provided the basis of all the subjects in the study; the other three groups were related to the first such that test-retest reliability and pre-test/post-test correlation could be included, and later manipulated, in the simulation.

The IRT ability of the second group was transformed such that it correlated to the IRT ability of the first group. The correlation of test scores between

Groups 1 and 2 provided the estimate of the test-retest reliability, and as expected, recovered the correlation of IRT ability, though with some reduction that is due (1) to the dichotomization of the item scores and (2) to the noise admitted by using the random number drawn from the uniform distribution.

The IRT ability of the third group was transformed such that it correlated to the IRT ability of the first. The correlation between Groups 2 and 3 total scores recovered the correlation between the IRT abilities, though with a reduction as seen between Groups 1 and 2, and provided the estimate of the correlation between pre- and post-test.

The IRT ability of the fourth group was transformed such that it correlated with the IRT ability of the third group. The correlation of the total scores between Groups 3 and 4 provided an estimate of the test-retest reliability of the post-test, and recovered, though with some slight reduction, the correlation of IRT abilities introduced between Groups 3 and 4.

These four groups provided two pair of matched subjects, Groups 1 & 3 and Groups 2 & 4, between which gain scores could be compared; then the pre-test reliability, the post-test reliability, the correlation between pre-test and post-test, and finally the change in variance from pre-test to post-test were manipulated to study the effects these have on the reliability of gain scores in the context of empirical data.

The following SAS System code was used to perform these simulations.

```
%macro reliable;
%do reps=1 %to 3 %by 1;
%do xx=1 %to 11 %by 1;
%do xy=1 %to 11 %by 1;

data diffs;
array a[100] a1-a100;
array b[100] b1-b100;
array c[100] c1-c100;
array hx[1000] hx1-hx1000;
array hxp[1000] hxp1-hxp1000;
array hy[1000] hy1-hy1000;
array hyp[1000] hyp1-hyp1000;

keep tsx tsxp tsy tsyp dif1 dif2;

/* create item parameter vectors */

seed1 = int(10000000*ranuni(0));
do i = 1 to 100 by 1;
call rannor(seed1,a[i]);
a[i] = 0.30*a[i] + 1.0;
if a[i] lt 0.10 then a[i] = 0.10;
end;

seed1 = int(10000000*ranuni(0));
do i = 1 to 100 by 1;
call rannor(seed1,b[i]);
b[i] = 1.0*b[i] + 0.0;
end;

seed1 = int(10000000*ranuni(0));
do i = 1 to 100 by 1;
call rannor(seed1,c[i]);
c[i] = 0.05*c[i] + 0.20;
if c[i] lt 0 then c[i] = 0;
end;

/* create correlated subject thetas */
rxx = (&xx-1)/10;
rxy = (&xy-1)/10;
ryy = rxx;

seed1 = int(10000000*ranuni(0));
do i = 1 to 1000 by 1;
call rannor(seed1,hx[i]);
end;
seed1 = int(10000000*ranuni(0));
do i = 1 to 1000 by 1;
call rannor(seed1,hxp[i]);
hxp[i] = rxx*hx[i] + sqrt(1-rxx*rxx)*hxp[i];
end;
seed1 = int(10000000*ranuni(0));
do i = 1 to 1000 by 1;
call rannor(seed1,hy[i]);
hy[i] = rxy*hx[i] + sqrt(1-rxy*rxy)*hy[i];
hy[i] = hy[i] + 0.25;
end;
seed1 = int(10000000*ranuni(0));
do i = 1 to 1000 by 1;
call rannor(seed1,hyp[i]);
hyp[i] = ryy*hy[i] + sqrt(1-ryy*ryy)*hyp[i];
end;

seed1 = int(10000000*ranuni(0));
do i = 1 to 1000 by 1;
tsx = 0;
tsxp = 0;
tsy = 0;
tsyp = 0;
do j = 1 to 100 by 1;
p=c[j]+(1-c[j])/
(1+exp(-1.7*a[j]*(hx[i]-b[j])));
call ranuni(seed1,noise);
if p gt noise then tsx = tsx+1;
p=c[j]+(1-c[j])/
(1+exp(-1.7*a[j]*(hxp[i]-b[j])));
call ranuni(seed1,noise);
if p gt noise then tsxp = tsxp+1;
p=c[j]+(1-c[j])/
(1+exp(-1.7*a[j]*(hy[i]-b[j])));
call ranuni(seed1,noise);
if p gt noise then tsy = tsy+1;
p=c[j]+(1-c[j])/
(1+exp(-1.7*a[j]*(hyp[i]-b[j])));
call ranuni(seed1,noise);
if p gt noise then tsyp = tsyp+1;
end;

dif1 = tsx - tsy;
dif2 = tsxp - tsyp;
output;
end;

run;

proc corr data=diffs nosimple pearson;
var dif1;
with dif2;
run;

%end;
%end;
%end;
%mend reliable;

filename rholst 'c:\papers\gainscor\reliable.lst';
filename rholog 'c:\papers\gainscor\reliable.log';
proc printto print=rholst log=rholog new;
```

```
run;
reliable;
proc printto;
run;
```

DESCRIPTION OF MONTE-CARLO STUDIES

STUDY 1

This study, using 1000 subjects in each of the four groups, provided the backdrop for those to follow. The reliability of gain scores was examined as a bivariate function of pre-test reliability and pre-test/post-test correlation. Post-test reliability was made equal to the pre-test reliability; the variances of the pre-test and post-test scores were also held equal. This simulation models the use of a standardized measure for both the pre- and post-test. This study, as those that follow, was replicated three times, little or no change was observed across the replications.

STUDIES 2A AND 2B

These two studies were the same as Study 1 except that in Study 2A the reliability of the post-test was one-third less than that of the pre-test. That is, if the reliability of the pre-test was 0.7, then the reliability of the post-test was 0.7 - 0.7/3, or about 0.47. In Study 2B, the reliability of the post-test is one-third greater than that of the pre-test, though it was computed slightly differently to prevent reliabilities in excess of 1. For example, if the pre-test reliability was 0.7, then the post-test reliability was 0.7 + (1-0.7)/3, or 0.8.

STUDIES 3A AND 3B

These two studies were based on Study 1, except that the variance of the post-test was allowed to change. In Study 3A, the variance of the pre-test IRT ability was held at about 1 while that of the post-test was reduced to about 0.75. In Study 3B, the variance of the post-test IRT ability was increased to about 1.25 while that of the pre-test was still held at about 1.

RESULTS OF THE STUDIES

STUDY 1

The results of Study 1, shown in Figure 1, are in keeping with the general trends shown in Table 2 of Zimmerman (1994) which based on the equation

$$\rho_{DD'} = \frac{\frac{1}{2}(\rho_{XX'} + \rho_{YY'}) - \rho(T_X T_Y) \sqrt{\rho_{XX'} \rho_{YY'}}}{1 - \rho(T_X T_Y) \sqrt{\rho_{XX'} \rho_{YY'}}} \quad (1)$$

developed to demonstrate the effect of changes in pre- and post-test reliability on gain score reliability. A closer examination of the correspondence between Zimmerman's Table 1 and Figure 1 shows that the gain score reliabilities estimated by the simulation are slightly below those computed from the preceding formula. This is due to the slight loss of information incurred when the IRT responses are dichotomized. However, the general trends are congruent. The gain score reliability is increased (1) by improving the reliability of the pre- and post-test, assumed equal in this simulation, and (2) decreasing the correlation between the pre- and post-test scores.

STUDY 2A AND 2B

The results of these two studies also conformed to the results of Zimmerman (1994) as expressed in Equation (1). That is, increasing the reliability of the post-test improved the gain score reliability even in the case where the reliability of the pre-test is zero, while a corresponding decrease in post-test reliability depressed that of the gain score. As well, it can be seen from the inspection of Equation (1) that it is the increase of either the pre-test reliability or the post-test reliability, or both, that increases the gain score reliability; not just the increase of the post-test reliability over that of the pre-test.

For instance, compare the instance in Study (1) where the reliability of the pre-test (and post-test) was approximately 0.7, the correlation between pre- and post-test was approximately 0.5, and the resulting gain score reliability was approximately 0.55 to the corresponding instance in Study (2B) where the reliability of the post-test only was raised to approximately 0.8 and the resulting gain score reliability increased slightly to approximately 0.6, an improvement of about 0.05. Now consider the graph in Figure (1) where the reliability of both pre- and post-tests is 0.80; the resultant gain score reliability is about 0.65, about twice the relative gain realized when only the post-test is manipulated, and confirming the discussion of Feldt (1995).

Lastly, it is evident from Studies (2A) and (2B) that increasing the reliability of the post-test decreases, but does not remove, the reduction in gain score reliability that is introduced by the correlation of the pre- and post-test; and as that correlation grows toward one, the gain score reliability falls substantially more slowly than in Study (1).

RESULTS OF STUDIES 3A AND 3B

The results of Studies (3A) and (3B) conformed to the trends set out in the discussion of Feldt (1995) where the reliability of gain scores in the setting of unequal variances for the pre- and post-tests was given in Lord (1963) by

$$\rho_{DD'} = \frac{\rho_{xx}(\sigma_x^2) + \rho_{yy}(\sigma_y^2) - 2\sigma_{xy}}{\sigma_x^2 + \sigma_y^2 - 2\sigma_{xy}} \quad (2)$$

and where σ_{xy} is the population covariance between the pre- and post-tests in that increasing the post-test variance over that of the pre-test generally improves gain score reliability. Also, the general trends established in Study (1) are again readily apparent in this study.

Studies (3A) and (3B) also indicate that unequal variances between pre- and post-tests serve to reduce the depression in gain score reliability that results from pre- and post-test correlation; however, unequal variance does not remove the trend shown in Figure (1), though there is evidence from the comparison of Figures (3A) and (3B) that increased variance, not just a change in variance, is important. For instance in the extreme case where pre- and post-test correlation approached unity, there is evidence in Figure 3B to suggest that the increase by 25 percent in post-test variance was sufficient to provide a gain score reliability that approached 0.50; and while 0.5 is not a particularly high reliability as measures go, it is quite respectable for a gain score reliability under the condition of highly correlated pre- and post-tests.

RESULTS OF FOLLOW-UP STUDY 4

The results of Studies (2B) and (3B) individually confirm that increasing post-test reliability and variance individually result in improved gain score reliability; a comparison of Figures (2B) and (3B) suggests that there may be a differential effect, indicating that an increase in reliability of the post-test has a greater effect on gain score reliability than does an increase in post-test variance. To consider the combined effect, a follow-up study was performed in which the post-test reliability was increased as it was in Study (2B) while the post-test variance was increased as it was in Study (3B). The results of this fourth study are presented in Figure (4). The trends established in Study (1) are repeated as expected; the general features of Study (2B) are reproduced; and, lastly, the general improvement in gain score reliability seen in Study (3B) over Study (1) is seen. As expected from Equation (2), the improvement in gain score seen Study (4) is generally a composite of the improvements seen in Studies (2B) and (3B). For instance, consider the example where the correlation between the pre- and post-test scores is about 0.5 while the reliability of the pre-test is 0.7. The improvement in gain scores seen in Studies (2B) and (3B) over Study (1) are 0.05 and 0.13, respectively. In Study (4), the improvement in gain score reliability over that seen in Study (1) is 0.15. However, it appears the Equation (2) consistently, if slightly, under-estimates the gain score reliability in Study (4). For instance consider again the case of an estimated pre- and post-test correlation of 0.5 and a pre-test reliability of 0.7, Equation (2) gives $\rho_{DD'}$ as approximately 0.54, a bit less than the value, 0.60, suggested in Figure (4).

There is also evidence presented in Figure (4) to suggest that the effect of the change in variance on the gain score reliability in the presence of highly correlated pre- and post-tests is lost when the reliability of the post-test is also increased. Given the overall "lifting" of gain score reliability seen when post-test reliability is improved, it does not seem a tragic loss. The effect of the increase in post-test variance to improve gain score reliability was only seen for highly correlated pre- and post-tests; its influence was far less for the lower correlations. Increasing the reliability of the post-test, however, resulted in higher gain score reliabilities for every other situation except where the pre- and post-test correlation was high.

IMPLICATIONS

The present study suggests that there are indeed conditions under which the use of gain scores in assessment is perfectly acceptable, these conditions being (1) good pre- and post-test reliability and (2) low pre- and post-test correlation; however, these conditions may be more difficult to obtain than may be expected, and may require some planning, if not a good deal of serendipity, on the part of the researcher. For instance, if a gain score reliability of at least 0.6 is deemed necessary and the pre- and post-test reliability is approximately 0.7, then the correlation between pre- and post-test should not exceed approximately 0.3, and probably should be less, to provide sufficiently high gain score reliability. This may be difficult to achieve in the case of a "weak" treatment, perhaps even in the case of a "strong" treatment. Given that (1) the pre-test and post-test should assess the same, perhaps single, cognitive dimension and that (2) the substantive difference in the pre- and post-test scores should be that the post-test reflects that the student has learned more along that one dimension, one would expect that the pre-test and post-test would share variance to a fair, perhaps high, degree, regardless of how much additional learning occurred during the treatment, resulting in a moderate to substantial correlation between pre- and post-test scores. Unfortunately, this is the very circumstance that is likely to lead to difference scores of poor reliability.

The present study also suggests that increasing post-test reliability has a differential impact on gain score reliability over increasing post-test variance, though it certainly does not hurt gain score reliability if both occur as discussed in Feldt (1995). However, if presented with the situation of being able to increase post-test reliability even if such results in reduced post-test variance, the practitioner should take that step, as the improvement in gain score reliability achieved by increasing the post-test reliability is substantially greater than the corresponding decrease in gain score reliability that results from the decrease in post-test variance. Of course, if the increase in post-test reliability introduces an increase in the correlation between pre- and post-test, then more study may be required; it could easily be the case that the improvement in gain score reliability introduced by the increase in post-test reliability is more than offset by the decrease in gain score reliability simultaneously introduced by the increase in pre- and post-test correlation.

EPILOG

The findings of this research, while Monte-Carlo in nature, are certainly applicable to "real world" situations where the consideration of the gain score is a necessary component of educational assessment. Indeed, the simulations allow the consideration of situations that may be difficult to achieve with real data, one such example being the case of the pre- and post-test correlation being moderate to low, and provide the measurement professional with the ability to anticipate the reliability of gain scores before actually committing to their use. As well, the effects of various factors on gain score reliability are presented simultaneously, allowing for the direct comparison of differences. In closing while this paper represents an empirical study of the psychometric properties of gain scores, and by no means represents a study of appropriate measures of gain, it is reasonable at this point to reflect upon the use of the gain score as a measure of change in general and learning in specific. Arguments for the importance of the assessment of change abound, (Lim, 1981; Ewell, 1984; Nuttall, 1986; and Astin, 1987) and arguments for the usefulness of gain scores as measures of educational improvement are numerous (Astin & Ewell, 1985; and McMillan, 1988). Indeed, if one assumes that the test scores are unbiased estimators of ability or knowledge, then the difference between the two should be an equally unbiased estimator of change regardless of the issues embedded within reliability, however, the

problematic nature of gain score reliability is well-documented (Cronbach & Furby, 1970; Overall & Woodard, 1975; Warren, 1984; and Pike, 1992), though perhaps overstated in Cronbach & Furby (1970, p. 78) who write: "There appears to be no need to use measures of change as dependent variables and no virtue in using them."

Perhaps the issue should turn from the reliability of gain scores to the validity of gain scores. One does not need to dig very deeply into the assessment literature, performance, authentic, portfolio, or otherwise, to find that single measures of change are rarely considered appropriate; composites built from multiple sources are preferred. For instance, candidates for licensure in teaching are often assessed on a variety of dimensions, some oblique, some orthogonal, all appropriate to the domain, generating individual scores along those single dimensions that lead to a presumably more stable composite score that is derived according to a well-established plan. However, it is probably unwise, as Bond (1995) discusses, to assume that a more complex score will be inherently more fair than the components; indeed, all the issues associated with the individual unidimensional score attend the multi-dimensional assessment. Test bias, differential validity, differential reliability have not been banished from the realm; rather, they are just more difficult to detect, and to that end, it is essential that properties of the components of composites are well-understood.

REFERENCES

- Astin, A. W. (1987). *Achieving Educational Excellence*. San Francisco: Jossey-Bass.
- Astin, A. W., & Ewell, P. T. (1985). The value added debate . . . continued. *AAHE Bulletin*, 37, 11-13.
- Baird, L. L. (1988). Value-added: Using student gains as yardsticks of learning. In C. Adelman (Ed.), *Performance and Judgement: Essays on Principles and Practice in the Assessment of College Student Learning* (pp. 205-216). Washington, DC: U.S. Government Printing Office.
- Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally.
- Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart, and Winston, Inc.
- Cronbach, L. J., & Furby, L. (1970). How should we measure "change" — or should we? *Psychological Bulletin*, 74, 68-70.
- Ewell, P. T. (1984). *The Self-Regarding Institution: Information for Excellence*. Boulder, CO: National Center for Higher Education Management Systems.
- Feldt, L. S. (1995). Estimation of the reliability of differences under revised reliabilities of component scores. *Journal of Educational Measurement*, 32, 295-301.
- Gulliksen, H. (1950). *Theory of mental tests*. New York: Wiley.
- Humphreys, L. G. (1991). The relation of power of statistical tests to range of talent: A correction and amplification. *Applied Psychological Measurement*, 15, 267.
- Humphreys, L. G. (1993). Further comments on reliability and power of significance tests. *Applied Psychological Measurement*, 17, 11-14.
- Kennedy, J. J., & Bush, A. J. (1985). *An Introduction to the Design and Analysis of Experiments in Behavioral Research*. New York: University Press of America.
- Lim, R. L. (1981). Measuring pretest-posttest performance changes. In R. A. Berk (Ed.), *Educational Evaluation Methodology: The State of the Art* (pp. 94-109). Baltimore: Johns Hopkins University Press.
- Lim, R. L., & Slinde, J. A. (1977). The determination of the significance of

change between pre- and posttesting periods. *Review of Educational Research*, 47, 121-150.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ: Erlbaum.

McMillan, J. H. (1988). Beyond value-added education: Improvement alone is not enough. *Journal of Higher Education*, 59, 564-579.

Nuttall, D. L. (1986). Problems in the measurement of change. In D. L. Nuttall (Ed.), *Assessing Educational Achievement* (pp. 153-167). London: Falmer Press.

O'Connor, E. F., Jr. (1972). Extending classical test theory to the measurement of change. *Review of Educational Research*, 42, 73-97.

Overall, J. E., & Woodard, J. A. (1975). Unreliability of difference scores: A paradox for the measurement of change. *Psychological Bulletin*, 82, 85-86.

Pike, G. R. (1992). Lies, damn lies, and statistics revisited: A comparison of three methods of representing change. *Research in Higher Education*, 33, 71-84.

Rogosa, D. R., & Willett, J. B. (1983). Demonstrating the reliability of the difference score in the measurement of change. *Journal of Educational Measurement*, 20, 335-343.

Roskam, E. E. (1976). Multivariate analysis of change and growth: Critical review and perspectives. In D. N. M. De Gruiter and L. J. T. van der Kamp (Eds.), *Advances in Psychological and Educational Measurement* (pp. 111-133). New York: John Wiley.

Willett, J. B. (1988). Questions and answers in the measurement of change. In E. Z. Rothkopf (Ed.), *Review of research in education* (Vol. 15, pp. 345-422). Washington, DC: American Educational Research Association.

Zimmerman, D. W. (1994). A note on the interpretation of formulas for the reliability of differences. *Journal of Educational Measurement*, 31, 143-147.

Zimmerman, D. W., & Williams, R. H. (1982). Gain scores in research can be highly reliable. *Journal of Educational Measurement*, 19, 149-154.

Zimmerman, D. W., Williams, R. H., & Zumbo, B. D. (1993). Reliability of measurement and power of significance tests based on differences. *Applied Psychological Measurement*, 17, 1-9.

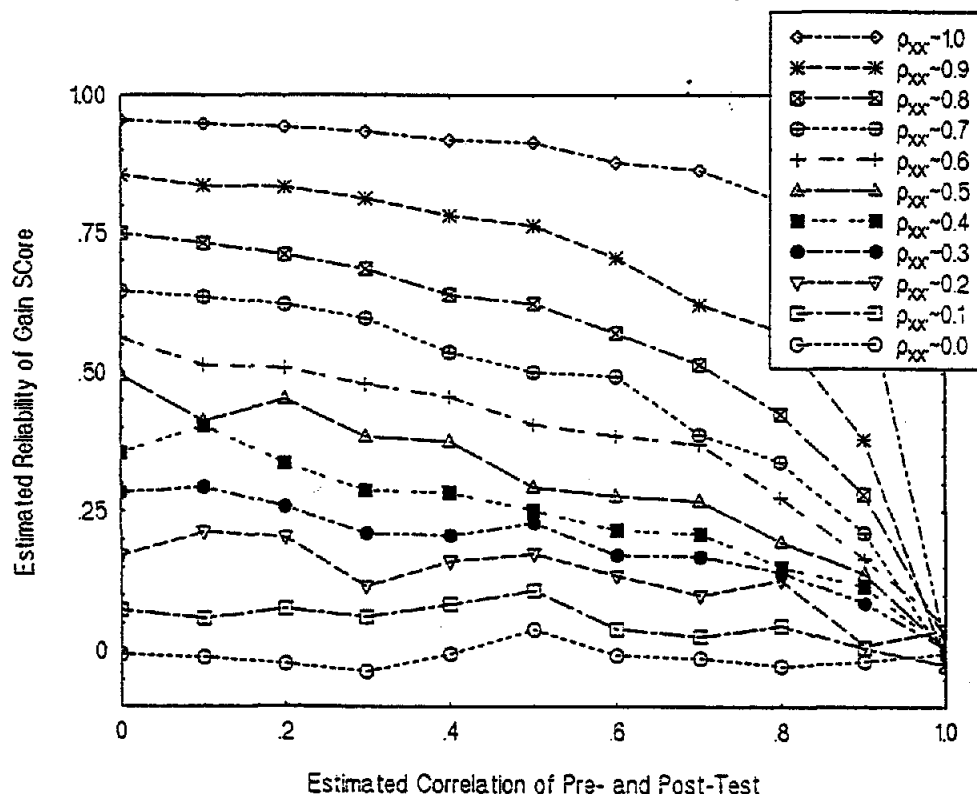


Figure 1
Effect of pre-test reliability and pre-test/post-test correlation on gain score reliability.

Notes

1. $\rho_{rr} = \rho_{xx}$
2. Pre-test and post-test have equal variances.

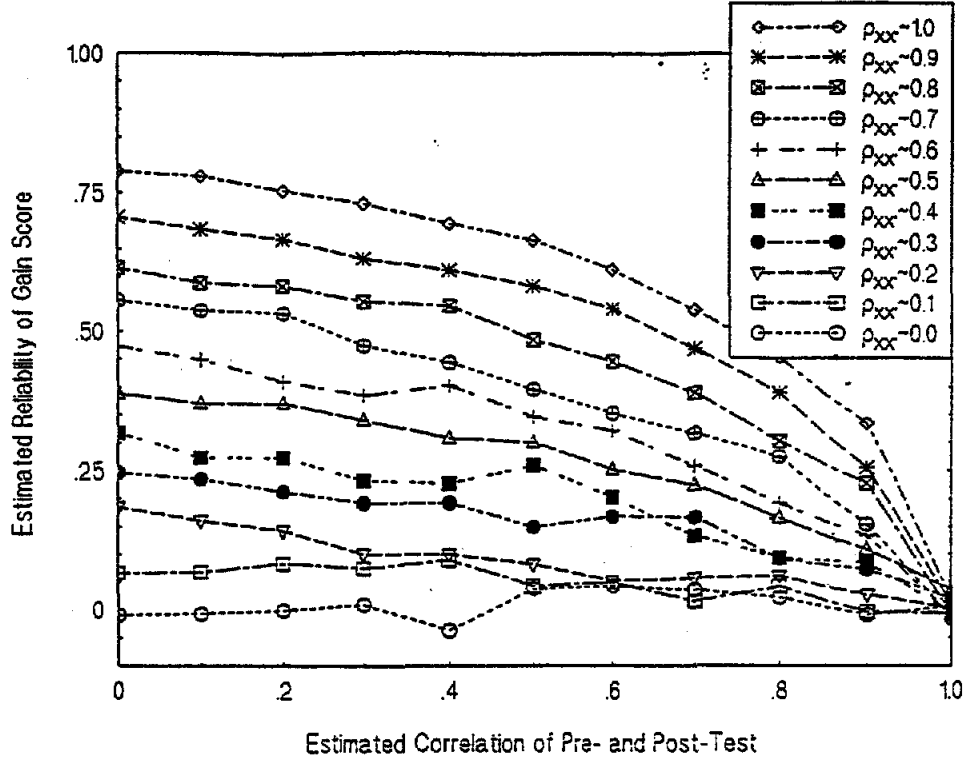


Figure 2a
Effect of pre-test reliability and pre-test/post-test correlation on gain score reliability.

Notes

1. $\rho_{rr} = \rho_{rr} - \rho_{rr}/3$
2. Pre-test and post-test have equal variances.

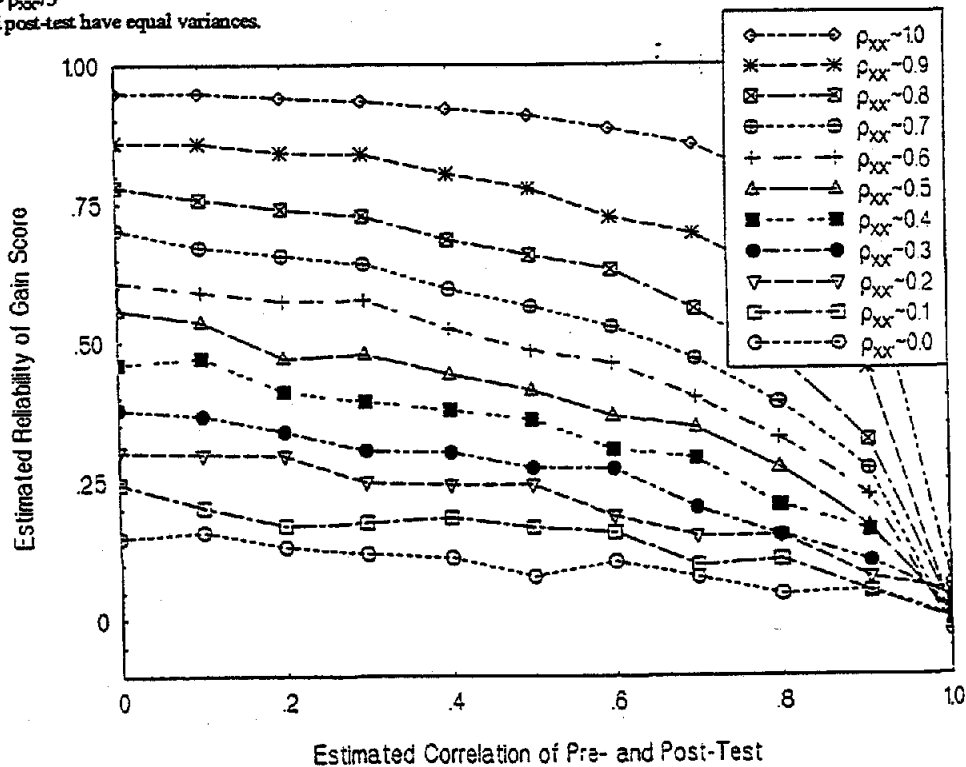


Figure 2b
Effect of pre-test reliability and pre-test/post-test correlation on gain score reliability.

Notes

1. $\rho_{rr} = \rho_{rr} + (1 - \rho_{rr})/3$
2. Pre-test and post-test have equal variances.

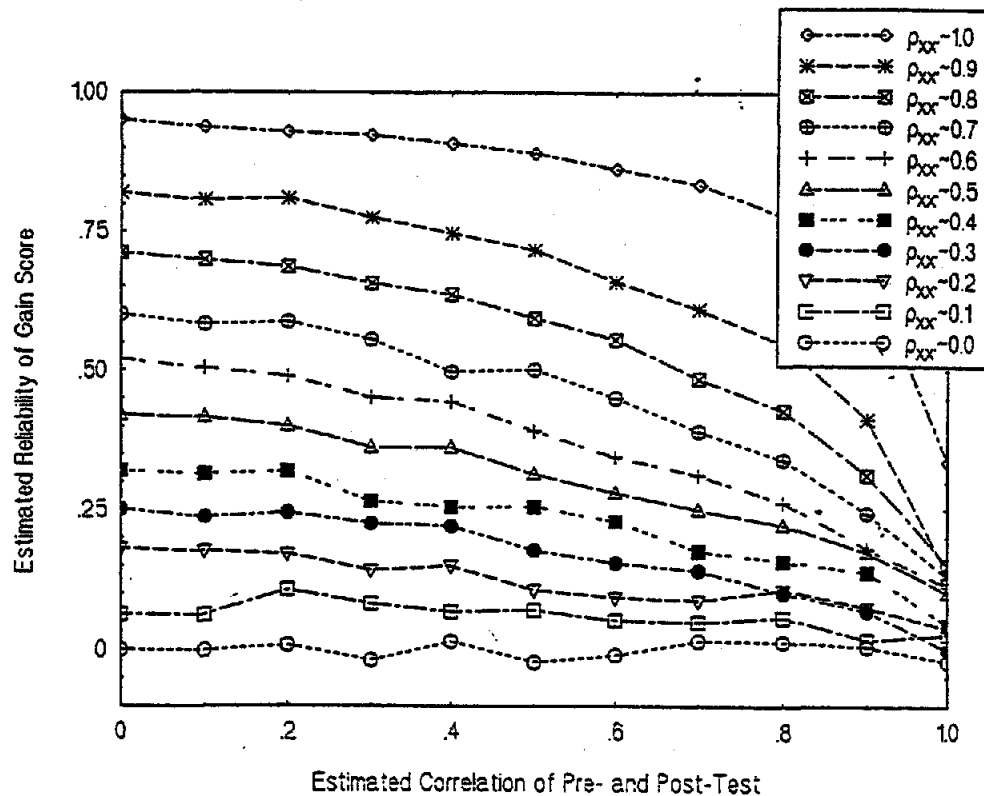


Figure 3a
Effect of pre-test reliability and pre-test/post-test correlation on gain score reliability.

Notes

1. $\rho_{yy} = \rho_{xx}$
2. $\sigma^2_{yy} = 0.75\sigma^2_{xx}$

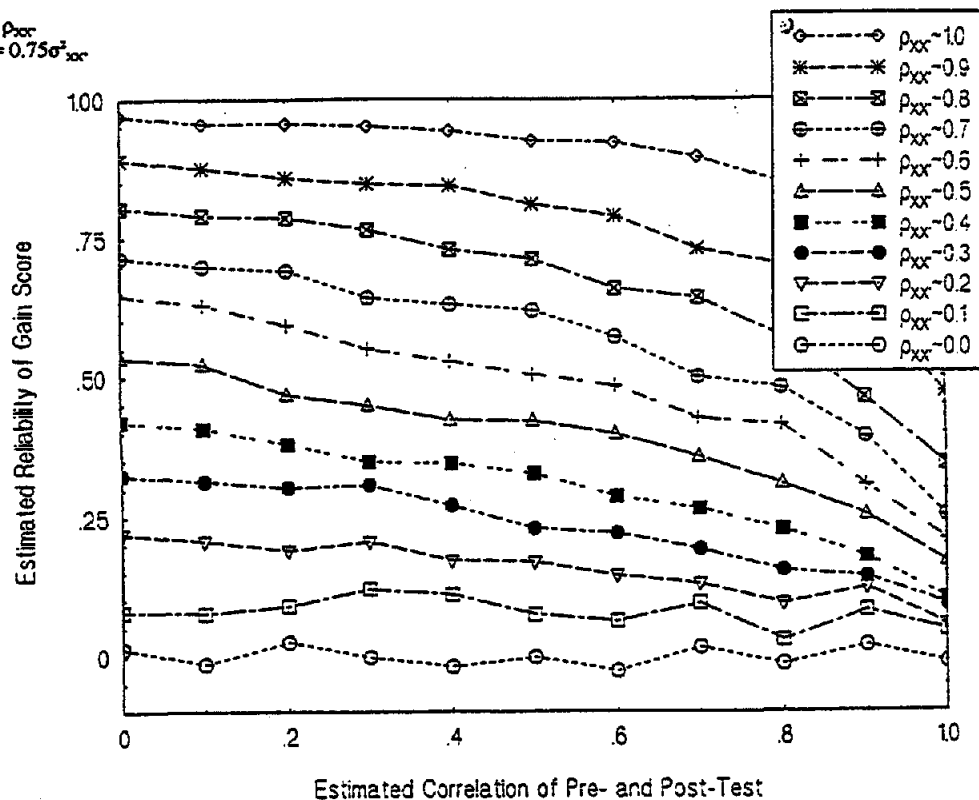


Figure 3b
Effect of pre-test reliability and pre-test/post-test correlation on gain score reliability.

Notes

1. $\rho_{yy} = \rho_{xx}$
2. $\sigma^2_{yy} = 1.25\sigma^2_{xx}$

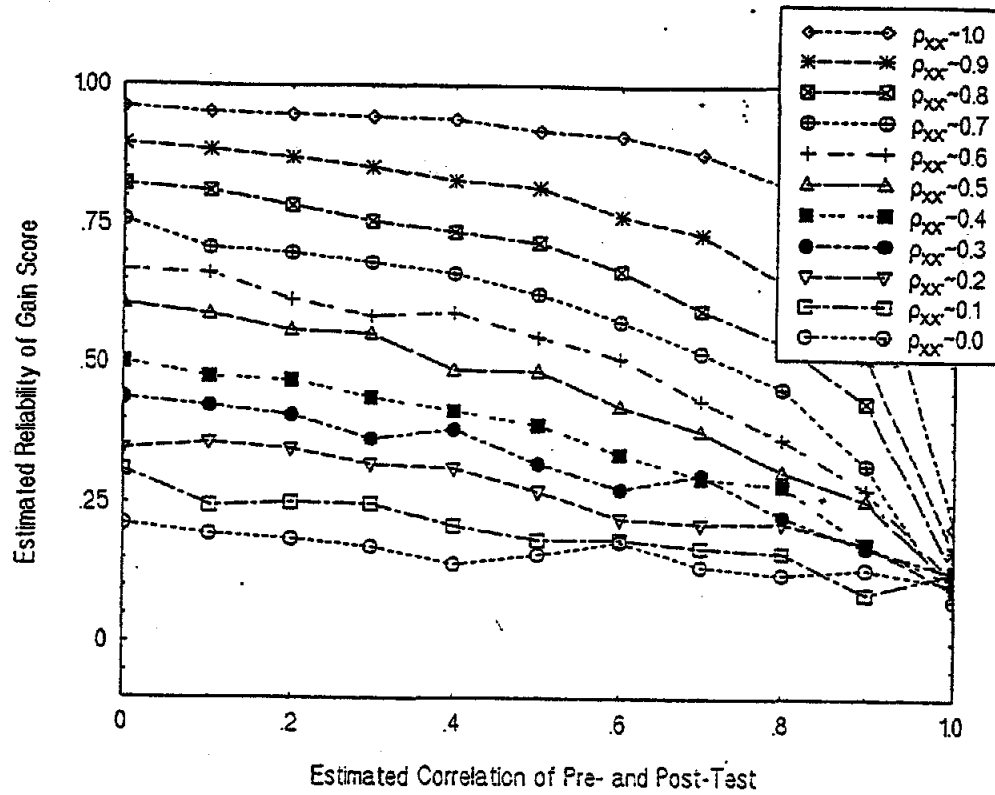


Figure 4
Effect of pre-test reliability and pre-test/post-test correlation on gain score reliability.

Notes

1. $\rho_{rr} = \rho_{xx} + (1 - \rho_{xx})/3$
2. $\sigma^2_{rr} = 1.25\sigma^2_{xx}$

SAS is a registered trademark of SAS Institute, Inc. in the USA and other countries.

Jim Penny
University Research Associates, Inc.
116 Tangle Drive
Jamestown, NC 27282
Voice and Fax: 910-454-4416
PENNYJ@IRIS.UNCG.EDU