DOCUMENT RESUME

ED 398 269 TM 025 398

AUTHOR Rachor, Robert E.; Cizek, Gregory J.
TITLE Reliability of Raw Gain, Residual Gain, and Estimated True Gain Scores: A Simulation Study.
PUB DATE Apr 96
NOTE 39p.; Paper presented at the Annual Meeting of the American Educational Research Association (New York, NY, April 8-12, 1996).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Achievement Gains; Correlation; Educational Research; Elementary Secondary Education; *Error of Measurement; Estimation (Mathematics); Measurement Techniques; *Pretests Posttests; *Raw Scores; Reliability; Simulation; Standardized Tests; *True Scores
IDENTIFIERS *Gain Scores; Residuals (Statistics)

ABSTRACT

The gain, or difference, score is defined as the difference between the posttest score and the pretest score for an individual. Gain scores appear to be a natural measure of growth for education and the social sciences, but they contain two sources of measurement error, error in either the pretest or posttest scores, and cannot be considered perfectly reliable. This study assessed the effect that different combinations of pretest and posttest reliability coefficients, pretest and posttest standard deviations, and the correlation between the pretest and posttest scores have on: (1) raw gain score reliability coefficients; (2) residual gain score coefficients; (3) estimated true score reliability coefficients; and (4) the correlation between pretest scores and raw gain scores. Simulations indicated that the reliability coefficients of the pretest and posttest scores determined to a large extent whether raw gain score, residual gain score, and estimated true gain scores were reliable. The study demonstrated that, for standardized tests of educational research such as those commonly encountered in elementary and secondary schools, gain scores can be useful indicators of progress if the user is knowledgeable about the proper type of gain score to use and its correct interpretation. (Contains 6 tables and 40 references.) (SLD)

Reliability of Raw Gain, Residual Gain, and Estimated True Gain Scores: A Simulation Study

Robert E. Rachor

Toledo Public Schools

and

Gregory J. Cizek

University of Toledo

A Paper Presented at the Annual Meeting of the

American Educational Research Association

New York, April, 1996

Correspondence about this paper should be directed to

Robert E. Rachor, Toledo Public Schools, 420 E. Manhattan Blvd,

Toledo, OH 43608, (419) 729-8434, e-mail RRACHOR@UOFT02.UTOLEDO.EDU

Background

The gain (or difference) score is defined as the difference between the posttest score and the pretest score for an individual. Gain scores appear to be a natural measure of growth for education and the social sciences. Willett (1988-1989) wrote that, "The very notion of learning implies growth and change" (p. 346). "Change phenomena ... such as the acquisition of knowledge, reduction of anxiety, positive changes in self-concept, and increase in productivity of human interactions are most validly viewed within the concept of change" (Corder-Bolz, 1978, p. 959).

Unfortunately, gain scores contain two sources of measurement error, the error in the pretest scores and the error in the posttest scores. Assuming that pre- and posttest scores are equally reliable, the two sources of error result in gain scores that are ordinarily less reliable than either the pre- or posttest scores. As the reliability of gain scores decreases, their usefulness in education and the social sciences also decreases. "Unreliability places a question mark after the score and causes any judgment based on it to be tentative to some extent. The accuracy of prediction that is possible to achieve is limited by the reliability of the measure through which the performance is being manifested" (Stanley, 1971, p. 358).

Historically, gain scores have been used for a variety of purposes. They have been used:

1) To represent the gain or loss of some skill for a specific individual (Rogosa, Brandt, & Zimowski, 1982);
2) As a dependent variable in an experimental or quasi-experimental research design (Cronbach & Furby, 1970; Fortune & Hutson, 1984; Rogosa, Brandt, & Zimowski, 1982);

3) As a criterion variable in a correlational study or a linear regression formula used in an attempt to predict future behavior (Cronbach & Furby, 1970; Fortune & Hutson, 1984; Rogosa, Brandt, & Zimowski, 1982);

4) To identify subjects for treatment or selection on the basis of their large or small gain scores (Cronbach & Furby, 1970; Fortune & Hutson, 1984); and,

5) To represent a construct such defining "self-satisfaction" as the "gain between ratings of self and ideal-self on an esteem scale" (Cronbach & Furby, 1970, p. 79).

Gain scores have both detractors and supporters with more of the former than the latter. Detractors recognize the intuitive appeal of gain scores, but assert that, "The fact that test scores are not perfectly reliable often makes this obvious procedure [the use of gain scores] produce absurd results" (Lord, 1956, p. 421).

Detractors list five reasons why gain scores are not appropriate. First, the posttest score is the most appropriate dependent variable in any experimental research design (Campbell & Stanley, 1966; Cronbach & Furby, 1970; Feldt, 1958; Knapp, 1980; Linn & Slinde, 1977). The advantage of pretest scores is that they can be used as a blocking variable or as a covariate to increase the precision of the analysis (Feldt, 1958).

Second, the use of gain scores in quasi-experimental designs may not prevent the confounding that exists between gain scores and pretest scores (Cronbach & Furby, 1970; Fortune & Hutson, 1984; Kenny, 1975; Linn & Slinde, 1977). When random assignment is not possible, the question must be asked: Are the gains or losses measured by the dependent variable (gain scores) the result of the treatment or pre-existing differences which existed before the treatment? (Kenny, 1975).

Third, it is well established that the correlation between a set of pretest scores and gain scores is ordinarily spuriously negative, even if there is no true correlation between the two variables (Bereiter, 1963; Cronbach & Furby, 1970; Diederich, 1956; Linn & Slinde, 1977; Lord, 1956, 1958, 1963; Thomson, 1924; Thorndike, 1924, 1966; Traub, 1994). Any group selected on the basis of their large raw gain scores will ordinarily contain an unusually large number of subjects with low pretest scores and an unusually small number of subjects with high pretest scores.

Fourth, raw gain scores have low reliability as derived by the procedures of classical test theory (Bereiter, 1963; Fortune & Hutson, 1984; Linn & Slinde, 1977; and Lord, 1963). Whenever the pre- and posttest reliability coefficients are equal and the pre- and posttest standard deviations are equal, raw gain score reliability coefficients are disappointingly low (Linn & Slinde, 1977).

Last, the pre- and posttest instruments may be measuring different constructs resulting in gain scores that are uninterpretable (Bereiter, 1963; Linn & Slinde, 1977; Lord, 1956, 1958; Traub, 1994). For example, parallel measures of the construct, mathematical ability may be measuring subtraction skills in the pretest measure (i.e. in a standardized achievement test at the end of second grade) and multiplication skills in the posttest measure (i.e. in a parallel standardized achievement test at the end of third grade); the result being a gain score that defies explanation. Lord (1958) cautioned that if the subjects under study have changed, even identical pre- and posttest instruments may not be measuring the same construct.

Supporters of gain scores (e.g. Engelhart, 1967; Maxwell & Howard, 1981; Overall & Woodward, 1975; Richards, 1975; Rogosa & Willett, 1983; and Zimmerman & Williams, 1982a, 1982b) admit that gain scores may not be reliable under most conditions. They insist, however, that in certain situations gain scores can be reliable; their point is that researchers should not automatically dismiss the use of gain scores when planning their research. Zimmerman and Williams (1982a) state that, "Our arguments indicate that gain scores can be reliable and it would be premature to discard such measures in research" (p. 153).

Some psychometricians point out that the assumption of equal pre- and posttest reliability coefficients and equal pre- and posttest standard deviations made by the detractors of gain scores is probably not realistic in longitudinal studies in education and the social sciences (Feldt & Brennan, 1989; Zimmerman & Williams, 1982a). Of particular interest to this study, Zimmerman and Williams assert that when the reliability coefficient and the standard deviation of the posttest scores exceeds the reliability coefficient and the standard deviation of the pretest scores, respectively, raw gain scores can be reliable.

Some critics of raw gain scores are less strongly opposed to--or even support--modified gain scores (Feldt & Brennan, 1989). One modified gain score is the residual gain score estimate. The residual gain score estimate is the difference between the actual and the predicted performance using a linear regression equation with the pretest score as the predictor variable.

Residual gain scores have an advantage over raw gain scores in that they are not correlated with the pretest score (Willett, 1988-1989). For this reason, residual gain scores may be useful to educators. One such use is the identification of individuals or schools where treatment (presumably, the educational program) has resulted in achievement gains which are greater or less than reasonably could be expected (Cronbach & Furby, 1970; O'Connor, 1972). The generally accepted belief is that residual gain scores are slightly more reliable than raw gain scores, particularly when pre- and posttest reliability coefficients are equal and pre- and posttest standard deviations are equal (DuBois, 1957; Linn & Slinde, 1977; Manning & DuBois, 1962; Veldman & Brophy, 1974).

A second alternative to the raw gain score approach is to estimate the true gain score. Whenever one talks about raw gain scores, the score of interest is the true gain score, that is, a gain score if there are no errors of measurement. One of the advantages of estimated true gain scores for research concerning student achievement is that estimated true gain scores are positively correlated with pretest scores. The result is that groups identified on the basis of their large gain scores contain many subjects with high pretest scores. This reflects the fact--not apparent when examining raw gain scores--that the brightest students tend to learn the most. When the pre- and posttest reliability coefficients are unequal and/or the pre- and posttest standard deviations are unequal, estimated true gain scores are considered more reliable than raw gain scores (Linn & Slinde, 1977).

Even though the prevailing wisdom among psychometricians appears to be that gain scores are unreliable and should be avoided as indicators of change, their use in evaluations and educational research is quite common. Zimmerman and Williams (1982a) have stated that, "Empirical studies are needed to determine how often and under what circumstances gain scores can be reliable in practical measurement situations" (p. 153).

Objectives

The factors affecting the reliability of raw gain scores and estimated true gain scores are the: 1) reliability coefficients of the pre- and posttest scores; 2) standard deviations of the pretest and posttest scores; and, 3) correlation between the pre- and posttest scores. The factors affecting the reliability coefficient of residual gain scores are: 1) the standard deviations of the pre- and posttest scores, and, 2) the correlation between the pre- and posttest scores. The factors affecting the correlation between raw gain scores and pretest scores are the; 1) reliability coefficients of the pre- and posttest scores; and, 2) the standard deviations of the pre- and posttest scores.

The purpose of this study was to investigate the effect that different combinations of the pre- and posttest reliability coefficients, pre- and posttest standard deviations and/or the correlation between the pre- and posttest scores have on: 1) raw gain score reliability coefficients; 2) residual gain score reliability coefficients, 3) estimated true gain score reliability coefficients, and, 4) the correlation between pretest scores and raw gain scores. The range of values for each of the three factors studied was limited, but included values appropriate to educational research concerning student achievement as measured by standardized achievement tests.

Method

This study began by identifying a reasonable range of values for the pre- and posttest reliability coefficients, standard deviations, and correlations between pre- and posttest scores to be used in this study. Of particular interest in this study were values of these statistics that would be of a magnitude commonly associated with those found on standardized measures of educational achievement used in elementary and secondary schools.

The following sections describe the values chosen for this study and details regarding the rationale for each.

The pre- and posttest reliability coefficients used for this study ranged from .75 to .95 in increments of .05, because the selected values covered the range of values reported in the technical manuals for the California Achievement Test, Form E and Iowa Test of Basic Skills. There seems to be general agreement that in longitudinal studies using different pre- and posttest measures, pre- and posttest reliability coefficients of stability and equivalence should be used in the calculation of gain score reliability coefficients (Feldt & Brennan, 1989; O'Connor, 1972; Stanley, 1967. The California Achievement Test, Form E & F Technical Manual (CTB/McGraw-Hill, 1987) reported that alternate form reliability coefficients for subtests in reading, arithmetic, and language arts, and for the total reading, total language, total arithmetic and total test battery for grades 3 through 12 ranged from .71 to .96. The Technical Manual (Psychological Corporation, 1993) reported that alternate form reliability coefficients for subtests in reading and for total reading, total mathematics, and total language tests ranged from .79 to .90.

Zimmerman and Williams (1982a, 1982b) algebraically manipulated the standard formula for the reliability of raw gain scores introducing lambda, the ratio of the standard deviation of the pretest scores to the standard deviation of the posttest scores. Their purpose was to illustrate the effect that the ratio of the two standard deviations had on raw gain score reliability coefficients. Similarly, formulas for the reliability of estimated true gain scores and the correlation between pretest scores and raw gain scores can be algebraically manipulated so that the effect of lambda on estimated true gain score reliability coefficients and the correlation between pretest scores and raw gain scores can be studied. The standard deviations of the pre- and posttest scores does not effect the reliability of residual gain scores.

To best illustrate the effect of lambda, the lambda values used in this study were .50, .67, .80, 1.00, 1.25, 1.50, and 2.00. The inverse of the lambda values of .50, .67, and .80 are the lambda values of 2.00, 1.50, and 1.25, respectively. Lambda values of 1.25, 1.50, and 2.00 were used to model situations when the standard deviation of the pretest scores was 25%, 50%, and 100% greater than the value for the standard deviation of the posttest scores. Lambda values of .50, .67, and .80 were used to model situations when the standard deviation of the posttest scores was 100%, 50%, and 25% greater than the standard deviation of the pretest scores.

Values selected for the correlation between the pre- and posttest scores for this study were .50 to .90 in increments of .10 because the selected values covered a plausible range of values. Two studies have used similar values. Martin (1985) reported in a study on raw gain scores that correlations between the pre- and posttest scores for the Iowa Test of Basic Skills (Riverside, 1978) ranged from .54 to .94. Rachor and Cizek (1994) in a study on the reliability of raw gain scores reported that the correlation between the pre- and posttest scores for the California Achievement Test, Form E (CTB/McGraw-Hill, 1985) ranged from .49 to .93.

There is a direct relationship between the values for the pre- and posttest reliability coefficients and the maximum possible value for the correlation between the pre- and posttest scores. The maximum value for the correlation between the pre- and posttest scores is less than or equal to the product of the square roots of the reliability coefficients of the pre- and posttest scores (Crocker & Algina, 1986). Table 1 lists the highest possible value for the correlation between the pre- and posttest scores for all possible combinations of the pre- and posttest reliability coefficients which were used in this study. When raw gain, residual gain, and estimated true gain score reliability coefficients, and the correlation between pretest scores and raw gain scores were calculated, the values used for the correlation between the pre- and posttest scores were those values which varied across the range of values selected for this study and which were equal to or less than the highest possible value for the correlation between the pre- and posttest scores.

Insert Table 1 About Here

Procedures

Microsoft Excel 5.0 (Microsoft Corporation, 1985-94) was used to calculate reliability coefficients for raw gain, residual gain, and estimated true gain scores, and the correlation between pretest scores and raw gain scores using all possible combinations of the range of values selected for this study. The formula used to calculate raw gain score reliability coefficients was the Zimmerman and Williams (1982a) formula:

$$\rho_{DD'} = \frac{\rho_{xx'}\lambda + \rho_{yy'}\lambda^{-1} - 2\rho_{xy}}{\lambda + \lambda^{-1} - 2\rho_{xy}} \qquad (1)$$

The reliability of the residual gain score was calculated using the Linn and Slinde (1977) formula:

$$\rho_{RR'} = \frac{\rho_{yy'} - \rho_{xy}^2\left(2 - \rho_{xx'}\right)}{1 - \rho_{xy}^2} \tag{2}$$

The reliability of estimated true gain scores was calculated using the Lord (1956, 1963) formula:

$$\rho_{\hat{G}G}^2 = \frac{\rho_{Gx}^2 + \rho_{Gy}^2 - 2\rho_{Gx}\rho_{Gy}\rho_{xy}}{1 - \rho_{xy}^2} \tag{3}$$

Lord supplied formulas for calculating the coefficients ρ_{Gx} and ρ_{Gy} used in Equation 3 (Lord, 1956, 1963). Lord's formulas can be algebraically manipulated to illustrate the effect of lambda on estimated true gain scores. The resulting formulas are:

$$\rho_{Gx} = \frac{\sqrt{\lambda^{-1}}\rho_{xy} - \sqrt{\lambda}\rho_{xx'}}{\sqrt{\rho_{DD'}\left(\lambda - 2\rho_{xy} + \lambda^{-1}\right)}} \tag{4}$$

$$\rho_{Gy} = \frac{\sqrt{\lambda^{-1}}\rho_{yy'} - \sqrt{\lambda}\rho_{xy}}{\sqrt{\rho_{DD'}\left(\lambda - 2\rho_{xy} + \lambda^{-1}\right)}} \qquad (5)$$

The standard formula for the correlation between the pretest scores and gain scores can be algebraically manipulated in a similar fashion so that the effect of lambda on the correlation between the pretest and gain scores can be demonstrated. The formula becomes:

$$\rho_{xg} = \frac{\sqrt{\lambda^{-1}}\rho_{xy} - \sqrt{\lambda}}{\sqrt{\left(\lambda - 2\rho_{xy} + \lambda^{-1}\right)}} \qquad (6)$$

Results

Tables 2 through 6 present raw gain score, residual gain score, and estimated true gain score reliability coefficients, and the correlation between the pretest scores and raw gain scores when pretest reliability coefficients were .75, .80, .85, .90, and .95, respectively, and the posttest reliability coefficients, lambda, and the correlation between the pre- and posttest scores varied across the limited range of values selected for this study.

Insert Tables 2 through 6 About Here

Raw Gain Score Reliability Coefficients

Raw gain score reliability coefficients increased as the pre- and/or posttest reliability coefficients increased. Regardless of the value of lambda or the correlation between the pre- and posttest scores, most raw gain score reliability coefficients were equal to or greater than .70 when one of the pre- or posttest reliability coefficients was at least .85 and the other was at least .90.

When pre- and posttest reliability coefficients were equal, increasing the pre- and posttest reliability coefficients by .05 resulted in increases in raw gain score reliability coefficients ranging from .08 to .50. The median increase was .12 and the majority of the increases fell in the .08 to .25 range. Regardless of the value for the pre- and posttest reliability coefficients, the increases in raw gain score reliability coefficients were relative consistent as pre- and posttest reliability coefficients increased by .05 while lambda and the correlation between the pre- and posttest scores were held constant. Increases in raw gain score reliability coefficients greater than .25 occurred when the pre- and posttest reliability coefficients increased from .85 to .90 or .90 to .95 and the correlation between the pre- and posttest scores was .80 or above. The large increases occurred because the raw gain score reliability coefficient for the score with the lower reliability coefficient was very low.

When only one of the pre- or posttest reliability coefficients increased by .05, raw gain score reliability coefficients increased from .01 to .25 with the majority of the increases in the .02 to .11 range. Larger increases tended to occur as the correlation between the pre- and posttest scores approached maximum possible values, again, because the raw gain score reliability coefficient for the score with the lower reliability coefficient was very low.

Increasing the correlation between the pre- and posttest scores resulted in decreases in raw gain score reliability coefficients. When the correlation between the pre- and posttest scores increased .10, decreases in raw gain score reliability coefficients ranged from .02 to .38 with the vast majority of the decreases ranging from .03 to .18. For specific values for the pre- and posttest reliability coefficients and lambda, larger decreases in raw gain score reliability coefficients were associated with increases in the correlation between the pre- and posttest scores from the second highest value used in this study to the maximum possible correlation. The larger decreases occurred because the raw gain score reliability coefficient for the score with the higher correlation between the pre- and posttest score was very low. Larger decreases also tended to occur as lambda approached one and smaller decreases tended to occur when lambda approached the largest or smallest values of lambda selected for this study. Larger decreases were also associated with lower values for the pre- and posttest reliability coefficients.

The effect of lambda on raw gain score reliability coefficients was dependent on the values for the pre- and posttest reliability coefficients. When pre- and posttest reliability coefficients were identical smaller values for raw gain score reliability coefficients occurred as lambda approached one; larger values for raw gain score reliability coefficients occurred as lambda diverged from one. When pretest reliability coefficients were larger than posttest reliability coefficients, raw gain score reliability coefficients increased as lambda increased from one; when the pretest reliability coefficients was .90 or above and lambda was 2.00, most raw gain score reliability coefficients were .70 or larger. Conversely, when pretest reliability coefficients were smaller than posttest reliability coefficients, raw gain score reliability coefficients increased as lambda decreased from one; when the posttest reliability coefficients were .90 or above and lambda was .50, most raw gain score reliability coefficients were .70 or larger.

Residual Gain Score Reliability Coefficients

Residual gain score reliability coefficients increased as the pre- and/or posttest reliability coefficients increased. Most residual gain score reliability coefficients were at least .70 when the posttest reliability coefficient was at least .85 and the pretest reliability coefficient was at least .90.

When the pre- and posttest reliability coefficients were equal, increasing the pre- and posttest reliability coefficients by .05, resulted in increases in residual gain score reliability coefficients ranging from .08 to .47; all but one of the increases were .23 or less. The median increase was .125. The increases in residual gain score reliability coefficients were similar for specific values for the correlation between the pre- and posttest scores; increasingly higher values for the correlation between the pre- and posttest scores were associated with larger increases in residual gain score reliability coefficients.

Increases in residual gain score reliability coefficients were more dependent on the posttest reliability coefficient than the pretest reliability coefficient. When pretest reliability coefficients were held constant and posttest reliability coefficients were increased by .05, residual gain score reliability coefficients increases ranged from .06 to .26, with all but two of the increases less than .15; The median increase was .08. Increases in residual gain score reliability coefficients were relatively consistent for fixed values of the correlation between the pre- and posttest scores.

When posttest reliability coefficients were held constant and pretest reliability coefficients were increased by .05, residual gain score reliability coefficient increases ranged from .01 to .21 with all but two of the increases were less than .10. The median increase was .03. Increases in residual gain score reliability coefficients were relatively consistent for fixed values of the correlation between the pre- and posttest scores.

Residual gain score reliability coefficients decreased as the correlation between pre- and posttest scores increased. Increasing the correlation between pre- and posttest scores by .10 resulted in decreases in residual gain score reliability coefficients ranging from .03 to .37. Most decreases were in the .06 to .24 range. The median decrease was .11. All of the decreases above .18 were associated with the largest possible value for the correlation between the pre- and posttest scores. For fixed values of the correlation between the pre- and posttest scores, differences in residual gain score reliability coefficients increased as the pre- or posttest reliability coefficients decreased.

Estimated True Gain Score Reliability Coefficients

Estimated true gain score reliability coefficients increased as the pre- and/or posttest reliability coefficients increased. Regardless of the value for lambda and the correlation between the pre- and posttest scores most estimated true gain score reliability coefficients were equal to or greater than .70 when pre- and posttest reliability coefficients were at least .90. When pre- and posttest reliability coefficients were equal and lambda was equal to one, estimated true gain score reliability coefficients were identical to raw gain score reliability coefficients.

When the correlation between the pre- and posttest scores was at its maximum possible value, estimated true gain score reliability coefficients increased dramatically, for all values of lambda which were different from one. For example, when the pre- and posttest reliability coefficients and the correlation between the pre- and posttest scores was .90, the estimated true gain score reliability coefficient was .81. However when the correlation between the pre- and posttest scores dropped to .895, .890, or .850, the estimated true gain score reliability coefficient dropped to .69, .61, or .52, respectively. Since it is unlikely that the correlation between pre- and posttest scores would reach the exact maximum possible value, the high estimated true gain score reliability coefficients have little practical value. Therefore, estimated true gain score reliability coefficients, when the correlation between the pre- and posttest scores was at its maximum possible value, will be ignored in this analysis.

When pre- and posttest reliability coefficients were equal, increasing the pre- and posttest reliability coefficients by .05 resulted in differences in estimated true gain score reliability coefficients ranging from decreases of .45 to increases of .54. The median increase in estimated true gain score reliability coefficients was .08. Most increases were in the .06 to .11 range. Larger increases in estimated true gain score reliability coefficients were associated with increasing values for the correlation between the pre- and posttest scores. All of the negative increases in estimated true gain score reliability coefficients were associated with values for the pre- and posttest reliability coefficients and the correlation between the pre- and posttest scores that were identical, thus producing very high estimated true gain score reliability coefficients as explained in the previous paragraph Regardless of the value of the pre- and posttest reliability coefficients, the increases in estimated true gain score reliability coefficients were relatively consistent as pre- and posttest reliability coefficients increased by .05 while lambda and the correlation between the pre- and posttest scores were held constant.

When only one of the pre- or posttest reliability coefficients was increased .05, differences between estimated true gain score reliability coefficients ranged from a decrease of .12 to an increase of .24. The median difference was .05. Most differences in estimated true gain score reliability coefficients ranged from -.02 to .10. Most decreases in estimated gain score reliability coefficients were associated with lambda values of .50, .67, 1.50, or 2.00. Most large increases in estimated true gain score reliability coefficients were associated with lambda values close to one or correlations between pre- and posttest scores that approached their maximum possible value.

Increasing the correlation between the pre- and posttest scores resulted in differential effects in estimated true gain score reliability coefficients. Changes in estimated true gain score reliability coefficients ranged from a decrease of .23 to an increase of .13 when differences associated with the maximum possible value for the correlation between the pre- and posttest scores were ignored. The median difference was -.05. Positive increases tended to occur when the correlation between the pre- and posttest scores approached their maximum possible value or when lambda values were less than or equal to .67 or more than or equal to 1.50.

When pretest reliability coefficients were larger than posttest reliability coefficients, estimated true gain score reliability coefficients increased as lambda values increased from one. When pretest reliability coefficients were smaller than posttest reliability coefficients, estimated true gain score reliability coefficients increased as lambda values decreased from one.

Correlation Between Pretest and Raw Gain Scores

The correlation between pretest scores and raw gain scores was primarily determined by the value of lambda. As lambda values decreased from .80 and the correlation between the pre- and posttest scores increased, the correlation between pretest scores and raw gain scores increased in value. Values for the correlation between pretest scores and raw gain scores were generally non-negative when lambda values were .67 or less. When lambda values were 1.00 or greater, the correlation between pretest scores and raw gain scores were negative across all values for the correlation between the pre- and posttest scores selected for this study. As lambda values increased from 1.25, maximum values for the correlation between pretest scores and raw gain scores were -.60.

Raw Gain Score Reliability Coefficients and the Correlation Between Pre- and Posttest Scores

The two major psychometric objections to the use of raw gain scores, as noted earlier in this paper, are the low reliability of raw gain scores and the spurious negative correlation between pretest scores and raw gain scores. When lambda values were .67 or less (i.e. when the posttest standard deviation was at least thirty-three percent larger than the pretest standard deviations) the correlation between raw gain scores and pretest scores were either positive values or small negative values. When lambda values were less than or equal to .67 and pre- and posttest reliability coefficients were at least .85, most raw gain score reliability coefficients were at least .70.

Comparison of Raw Gain, Residual Gain, and
Estimated True Gain Score Reliability Coefficients and
the Correlation Between Pre- and Posttest Scores

In situations where lambda was equal to one, residual gain score reliability coefficients are higher than raw gain or estimated true gain score reliability coefficients. Under these constraints, raw gain scores are not appropriate due to the negative correlation between raw gain and pretest scores. Residual gain scores would seem to be preferable under these constraints, particularly when pre- and posttest reliability coefficients were less than .90. At values at or above .90, the differences between residual gain and estimated true gain scores were minimal. In practice, this means that residual gain scores are most likely to be preferable--when the pre- and posttest score distributions can be expected to have equal variability.

Estimated true gain score reliability coefficients were higher than raw gain or estimated true gain score reliability coefficients under several circumstances. When the correlation between pre- and posttest scores was close to its maximum possible value, estimated true gain score reliability coefficients were always higher than raw gain or residual gain scores (other than the circumstance when lambda was equal to one).

When treatment markedly affect the standard deviation of the posttest scores (and thus lambda), estimated true gain scores may also be preferable. When lambda values were .50 or 2.00, estimated true gain score reliability coefficients were higher than raw gain or residual gain score reliability coefficients regardless of the value for the pre- or posttest reliability coefficients or the correlation between the pre- and posttest scores. When lambda values were .67 or less or 1.50 or more, values for estimated true gain score reliability coefficients were quite consistent across all values for the correlation between the pre- and posttest scores, while raw gain and residual gain score reliability coefficients were much more affected by the correlation between the pre- and posttest scores; in most cases estimated true gain score reliability coefficients were higher than raw gain or residual gain score reliability coefficients. Thus estimated true gain scores would seem to be preferable when the correlation between pre- and posttest scores would likely be quite high, such as in studies of a short duration or when little change in the rank of students on the dependent variable is unlikely to change or when the pre- or posttest standard deviation is at least 33% larger than the post- or pretest standard deviation, respectively.

In many cases, the differences between raw gain score, residual gain score and/or estimated true gain score reliability coefficients were minimal. Raw gain scores would seem to be preferable because of the laws of parsimony in circumstances where lambda is less than .80 and the correlation between raw gain scores and pretest scores is non-negative,.

Summary

The purpose of this study was to investigate the effect that different combinations of the pre- and posttest reliability coefficients, lambda, and/or the correlation between the pre- and posttest scores had on the reliability of raw gain, residual gain, and/or estimated true gain scores and the correlation between raw gain scores and pretest scores. The reliability coefficients of the pre- and posttest scores determined to a large extent whether raw gain score, residual gain score, and estimated true gain scores were reliable. Some specific findings include:

- Lambda values of one or greater than one were associated with negative correlations between pretest scores and raw gain scores, thus discouraging the use of raw gain scores in that situation.

- Residual gain scores are generally most reliable when lambda values were around one.
- Estimated true gain scores seem to be preferred in situations when lambda values were more extreme or when the correlation between the pre- and posttest scores approached the maximum possible value.
- The principle of parsimony argues for the use of raw gain scores when reliability coefficients of the three types of gain scores are similar and the correlation between the pretest scores and raw gain scores is close to zero or is positive.

Schools are often evaluated on their ability to increase the achievement of their students. When communicating about student progress and evaluating school programs, most principals and teachers talk about the gains of students. There is also a recognized need for reliable gain scores in quasi-experimental research and evaluation designs. In educational and social science research there are many situations where random selection of subjects for treatment and control groups is not possible. However, substantive questions remain about the effectiveness of particular treatments within these settings.

This study has demonstrated that, for standardized tests of educational research such as those commonly encountered in elementary and secondary schools, gain scores can be useful indicators of progress, <u>if the user is knowledgeable about the proper type of gain score to use and knowledgeable about its proper interpretation.</u> This study has investigated raw gain scores, residual gain scores, and estimated true gain scores and has generated some guidance regarding the appropriate matching of these tools to the educational/measurement contexts in which their use is most likely to yield defensible, interpretable results. A fruitful line of inquiry for the future may be investigation of the utility of these tools in other measurement contexts.

Table 1

Maximum Values for the Correlation Between the Pre- and Posttest Scores

Pre- and Posttest Reliability Coefficients					
	.75	.80	.85	.90	.95
.75	.75				
.80	.77	.80			
.85	.79	.82	.85		
.90	.82	.84	.87	.90	
.95	.84	.87	.89	.92	.95

Table 2
Raw Gain Score, Residual Gain Score, and Estimated True Gain Score Reliability Coefficients Assuming $\rho_{xx'}$=.75

	λ=.50			λ=.67			λ=.80			λ=1.00			λ=1.25			λ=1.50			λ=2.00			
ρ_{xy}	R	Est	ρ_{xg}	R	Est	ρ_{xg}	R	Est	ρ_{xg}	R	Est	ρ_{xg}	R	Est	ρ_{xg}	R	Est	ρ_{xg}	R	Est	ρ_{xg}	Re
$\rho_{yy'}$=.75																						
.50	.58	.62	.00	.53	.55	-.19	.51	.52	-.33	.50	.50	-.50	.51	.52	-.65	.54	.56	-.76	.58	.62	-.87	.58
.60	.52	.61	.12	.44	.50	-.09	.40	.42	-.24	.38	.38	-.45	.40	.42	-.63	.44	.50	-.75	.52	.61	-.87	.47
.70	.43	.69	.27	.29	.53	.04	.21	.35	-.14	.17	.17	-.39	.21	.35	-.61	.29	.54	-.75	.43	.69	-.88	.27
.80	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
.90	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
$\rho_{yy'}$=.80																						
.50	.65	.68	.00	.60	.62	-.19	.57	.58	-.33	.55	.55	-.50	.55	.55	-.65	.56	.57	-.76	.60	.62	-.87	.65
.60	.60	.67	.12	.52	.57	-.09	.47	.50	-.24	.44	.44	-.45	.44	.45	-.63	.47	.51	-.75	.54	.60	-.87	.55
.70	.52	.71	.27	.39	.56	.04	.31	.42	-.14	.25	.25	-.39	.27	.33	-.61	.34	.48	-.75	.45	.64	-.88	.37
.80	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
.90	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
$\rho_{yy'}$=.85																						
.50	.72	.75	.00	.66	.69	-.19	.63	.65	-.33	.60	.65	-.50	.59	.59	-.65	.59	.60	-.76	.62	.63	-.87	.72
.60	.67	.73	.12	.59	.64	-.09	.54	.58	-.24	.50	.51	-.45	.49	.49	-.63	.51	.52	-.75	.56	.60	-.87	.63
.70	.61	.74	.27	.49	.62	.04	.40	.50	-.14	.33	.35	-.39	.33	.35	-.61	.38	.46	-.75	.48	.61	-.88	.47
.80	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
.90	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
$\rho_{yy'}$=.90																						
.50	.78	.81	.00	.73	.75	-.19	.69	.71	-.33	.65	.66	-.50	.63	.63	-.65	.62	.62	-.76	.63	.64	-.87	.78
.60	.75	.80	.12	.67	.72	-.09	.62	.66	-.24	.56	.58	-.45	.54	.54	-.63	.54	.55	-.75	.58	.60	-.87	.70
.70	.70	.80	.27	.58	.69	.04	.50	.59	-.14	.42	.44	-.39	.40	.40	-.61	.42	.46	-.75	.50	.59	-.88	.56
.80	.64	.88	.45	.44	.81	.21	.28	.68	.00	.13	.25	-.32	.13	.27	-.60	.22	.59	-.76	.39	.78	-.89	.28
.90	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
$\rho_{yy'}$=.95																						
.50	.85	.87	.00	.79	.82	-.19	.75	.78	-.33	.70	.72	-.50	.66	.67	-.65	.65	.65	-.76	.65	.65	-.87	.85
.60	.83	.87	.12	.75	.80	-.09	.69	.73	-.24	.63	.65	-.45	.59	.59	-.63	.58	.58	-.75	.60	.61	-.87	.78
.70	.80	.86	.27	.68	.77	.04	.60	.68	-.14	.50	.54	-.39	.46	.46	-.61	.47	.48	-.75	.52	.58	-.88	.66
.80	.75	.90	.45	.57	.81	.21	.42	.68	.00	.25	.36	-.32	.22	.24	-.60	.28	.45	-.76	.42	.68	-.89	.42
.90	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*

R - Raw Gain Score Reliability Coefficients
Re - Residual Gain Score Reliability Coefficients
Est - Estimated True Gain Score Reliability Coefficients
ρ_{xg} - Correlation between pretest scores and raw gain scores
* - Cannot be calculated since ρ_{xy} exceeds maximum possible value

Table 3

Raw Gain Score, Residual Gain Score, and Estimated True Gain Score Reliability Coefficients Assuming $\rho_{xx'}=.80$

	λ=.50			λ=.67			λ=.80			λ=1.00			λ=1.25			λ=1.50			λ=2.00			
ρ_{xy}	R	Est	ρ_{xg}	R	Est	ρ_{xg}	R	Est	ρ_{xg}	R	Est	ρ_{xg}	R	Est	ρ_{xg}	R	Est	ρ_{xg}	R	Est	ρ_{xg}	Re
$\rho_{yy'}=.75$																						
.50	.60	.62	.00	.56	.57	-.19	.55	.55	-.33	.55	.55	-.50	.57	.58	-.65	.60	.62	-.76	.65	.68	-.87	.60
.60	.54	.60	.12	.47	.50	-.09	.44	.45	-.24	.44	.44	-.45	.47	.50	-.63	.52	.57	-.75	.60	.67	-.87	.50
.70	.45	.64	.27	.33	.47	.04	.27	.33	-.14	.25	.25	-.39	.31	.42	-.61	.39	.57	-.75	.52	.71	-.88	.32
.80	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
.90	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
$\rho_{yy'}=.80$																						
.50	.67	.69	.00	.63	.64	-.19	.61	.61	-.33	.60	.60	-.50	.61	.61	-.65	.63	.64	-.76	.67	.69	-.87	.67
.60	.62	.66	.12	.55	.58	-.09	.52	.53	-.24	.50	.50	-.45	.52	.53	-.63	.55	.58	-.75	.62	.66	-.87	.58
.70	.55	.68	.27	.43	.54	.04	.37	.42	-.14	.33	.33	-.39	.37	.42	-.61	.43	.54	-.75	.55	.68	-.88	.42
.80	.44	.89	.45	.23	.89	.21	.09	.89	.00	.00	--	-.32	.09	.89	-.60	.24	.89	-.76	.44	.89	-.89	.09
.90	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
$\rho_{yy'}=.85$																						
.50	.73	.75	.00	.69	.71	-.19	.67	.68	-.33	.65	.65	-.50	.65	.65	-.65	.66	.66	-.76	.68	.69	-.87	.73
.60	.69	.73	.12	.63	.66	-.09	.59	.61	-.24	.56	.56	-.45	.56	.57	-.63	.59	.60	-.75	.63	.67	-.87	.65
.70	.64	.73	.27	.53	.61	.04	.47	.51	-.14	.42	.42	-.39	.43	.45	-.61	.48	.53	-.75	.57	.66	-.88	.51
.80	.56	.83	.45	.36	.72	.21	.23	.54	.00	.13	.14	-.32	.18	.40	-.60	.29	.65	-.76	.47	.79	-.89	.23
.90	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
$\rho_{yy'}=.90$																						
.50	.80	.82	.00	.76	.77	-.19	.73	.74	-.33	.70	.70	-.50	.69	.69	-.65	.69	.69	-.76	.70	.70	-.87	.80
.60	.77	.80	.12	.71	.73	-.09	.66	.69	-.24	.63	.63	-.45	.61	.61	-.63	.62	.63	-.75	.65	.67	-.87	.73
.70	.73	.79	.27	.63	.69	.04	.56	.61	-.14	.50	.51	-.39	.49	.50	-.61	.52	.55	-.75	.59	.65	-.88	.61
.80	.67	.83	.45	.50	.72	.21	.37	.56	.00	.25	.28	-.32	.27	.33	-.60	.35	.54	-.76	.50	.72	-.89	.37
.90	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
$\rho_{yy'}=.95$																						
.50	.87	.88	.00	.82	.84	-.19	.79	.80	-.33	.75	.75	-.50	.72	.73	-.65	.71	.71	-.76	.72	.72	-.87	.87
.60	.85	.87	.12	.78	.81	-.09	.74	.76	-.24	.69	.70	-.45	.66	.66	-.63	.66	.66	-.75	.67	.68	-.87	.81
.70	.82	.86	.27	.73	.78	.04	.66	.70	-.14	.58	.60	-.39	.55	.55	-.61	.57	.57	-.75	.61	.65	-.88	.71
.80	.78	.88	.45	.63	.78	.21	.51	.65	.00	.38	.42	-.32	.36	.37	-.60	.41	.50	-.76	.53	.67	-.89	.51
.90	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*

R - Raw Gain Score Reliability Coefficients
Re - Residual Gain Score Reliability Coefficients
Est - Estimated True Gain Score Reliability Coefficients
ρ_{xg} - Correlation between pretest scores and raw gain scores
* - Cannot be calculated since ρ_{xy} exceeds maximum possible value

Table 4

Raw Gain Score, Residual Gain Score, and Estimated True Gain Score Reliability Coefficients Assuming $\rho_{xx'}$=.85.

ρ_{xy}	λ=.50			λ=.67			λ=.80			λ=1.00			λ=1.25			λ=1.50			λ=2.00			
	R	Est	ρ_{xg}	R	Est	ρ_{xg}	R	Est	ρ_{xg}	R	Est	ρ_{xg}	R	Est	ρ_{xg}	R	Est	ρ_{xg}	R	Est	ρ_{xg}	Re
$\rho_{yy'}$=.75																						
.50	.62	.63	.00	.59	.59	-.19	.59	.59	-.33	.60	.61	-.50	.63	.65	-.65	.66	.69	-.76	.72	.75	-.87	.62
.60	.56	.60	.12	.51	.52	-.09	.49	.49	-.24	.50	.51	-.45	.54	.58	-.63	.59	.65	-.75	.67	.73	-.87	.53
.70	.48	.61	.27	.38	.45	.04	.33	.35	-.14	.33	.35	-.39	.40	.50	-.61	.49	.62	-.75	.61	.74	-.88	.37
.80	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
.90	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
$\rho_{yy'}$=.80																						
.50	.68	.69	.00	.66	.66	-.19	.65	.65	-.33	.65	.65	-.50	.67	.68	-.65	.69	.71	-.76	.73	.75	-.87	.68
.60	.63	.67	.12	.59	.60	-.09	.56	.57	-.24	.56	.56	-.45	.59	.61	-.63	.63	.66	-.75	.69	.73	-.87	.60
.70	.57	.66	.27	.48	.53	.04	.43	.45	-.14	.42	.42	-.39	.57	.51	-.61	.53	.61	-.75	.64	.73	-.88	.46
.80	.47	.79	.45	.29	.65	.21	.18	.40	.00	.13	.14	-.32	.23	.54	-.60	.37	.73	-.76	.56	.83	-.89	.18
.90	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
$\rho_{yy'}$=.85																						
.50	.75	.76	.00	.72	.73	-.19	.71	.71	-.33	.70	.65	-.50	.71	.71	-.65	.72	.73	-.76	.75	.76	-.87	.75
.60	.71	.74	.12	.66	.68	-.09	.64	.64	-.24	.63	.63	-.45	.64	.64	-.63	.66	.68	-.75	.71	.74	-.87	.68
.70	.66	.72	.27	.57	.62	.04	.53	.55	-.14	.50	.50	-.39	.53	.55	-.61	.58	.62	-.75	.66	.72	-.88	.56
.80	.58	.77	.45	.42	.63	.21	.32	.44	.00	.25	.25	-.32	.32	.44	-.60	.43	.63	-.76	.58	.77	-.89	.32
.90	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
$\rho_{yy'}$=.90																						
.50	.82	.83	.00	.79	.79	-.19	.77	.77	-.33	.75	.75	-.50	.75	.75	-.65	.75	.75	-.76	.77	.77	-.87	.82
.60	.79	.81	.12	.74	.75	-.09	.71	.72	-.24	.69	.69	-.45	.69	.69	-.63	.70	.70	-.75	.73	.74	-.87	.76
.70	.75	.79	.27	.67	.71	.04	.62	.64	-.14	.58	.59	-.39	.59	.59	-.61	.62	.64	-.75	.68	.72	-.88	.66
.80	.69	.81	.45	.56	.68	.21	.46	.54	.00	.38	.38	-.32	.41	.45	-.60	.49	.59	-.76	.61	.74	-.89	.46
.90	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
$\rho_{yy'}$=.95																						
.50	.88	.89	.00	.85	.86	-.19	.83	.83	-.33	.80	.80	-.50	.78	.78	-.65	.78	.78	-.76	.78	.78	-.87	.88
.60	.87	.88	.12	.82	.83	-.09	.79	.80	-.24	.75	.76	-.45	.73	.73	-.63	.73	.73	-.75	.75	.76	-.87	.84
.70	.84	.87	.27	.77	.80	.04	.72	.74	-.14	.67	.67	-.39	.65	.65	-.61	.66	.67	-.75	.70	.72	-.88	.76
.80	.81	.87	.45	.69	.77	.21	.59	.66	.00	.50	.51	-.32	.49	.51	-.60	.54	.59	-.76	.64	.72	-.89	.59
.90	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*

R - Raw Gain Score Reliability Coefficients
Re - Residual Gain Score Reliability Coefficients
Est - Estimated True Gain Score Reliability Coefficients
ρ_{xg} - Correlation between pretest scores and raw gain scores
* - Cannot be calculated since ρ_{xy} exceeds maximum possible value

Table 5

Raw Gain Score, Residual Gain Score, and Estimated True Gain Score Reliability Coefficients Assuming $\rho_{xx'}$=.90.

λ	λ=.50			λ=.67			λ=.80			λ=1.00			λ=1.25			λ=1.50			λ=2.00			
ρ_{xy}	R	Est	ρ_{xg}	R	Est	ρ_{xg}	R	Est	ρ_{xg}	R	Est	ρ_{xg}	R	Est	ρ_{xg}	R	Est	ρ_{xg}	R	Est	ρ_{xg}	Re
$\rho_{yy'}$=.75																						
.50	.63	.64	.00	.62	.62	-.19	.63	.63	-.33	.65	.66	-.50	.69	.71	-.65	.73	.76	-.76	.78	.81	-.87	.63
.60	.58	.60	.12	.54	.55	-.09	.54	.54	-.24	.56	.58	-.45	.62	.66	-.63	.67	.72	-.75	.75	.80	-.87	.55
.70	.50	.59	.27	.42	.46	.04	.40	.40	-.14	.42	.44	-.39	.50	.59	-.61	.59	.70	-.75	.70	.80	-.88	.41
.80	*	.78	*	*	.58	*	*	.27	*	*	.25	*	*	.68	*	*	.81	*	*	.88	*	*
.90	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
$\rho_{yy'}$=.80																						
.50	.70	.70	.00	.69	.69	-.19	.69	.69	-.33	.70	.70	-.50	.73	.74	-.65	.76	.80	-.76	.80	.82	-.87	.70
.60	.65	.67	.12	.62	.62	-.09	.61	.61	-.24	.63	.63	-.45	.66	.69	-.63	.71	.77	-.75	.77	.80	-.87	.63
.70	.59	.65	.27	.52	.55	.04	.49	.50	-.14	.50	.51	-.39	.56	.61	-.61	.63	.75	-.75	.73	.79	-.88	.51
.80	.50	.72	.45	.35	.53	.21	.27	.33	.00	.25	.28	-.32	.37	.56	-.60	.50	.80	-.76	.67	.83	-.89	.27
.90	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
$\rho_{yy'}$=.85																						
.50	.77	.77	.00	.75	.75	-.19	.75	.75	-.33	.75	.75	-.50	.77	.77	-.65	.79	.79	-.76	.82	.83	-.87	.77
.60	.73	.74	.12	.70	.70	-.09	.69	.69	-.24	.69	.69	-.45	.71	.72	-.63	.74	.76	-.75	.79	.81	-.87	.71
.70	.68	.72	.27	.62	.64	.04	.59	.59	-.14	.58	.59	-.39	.62	.64	-.61	.67	.71	-.75	.75	.79	-.88	.61
.80	.61	.74	.45	.48	.59	.21	.41	.45	.00	.38	.38	-.32	.46	.54	-.60	.56	.68	-.76	.69	.81	-.89	.41
.90	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
$\rho_{yy'}$=.90																						
.50	.83	.84	.00	.81	.82	-.19	.80	.81	-.33	.80	.80	-.50	.80	.81	-.65	.81	.82	-.76	.83	.84	-.87	.83
.60	.81	.82	.12	.78	.78	-.09	.76	.76	-.24	.75	.75	-.45	.76	.76	-.63	.78	.78	-.75	.81	.82	-.87	.79
.70	.77	.80	.27	.72	.73	.04	.68	.69	-.14	.67	.67	-.39	.68	.69	-.61	.72	.73	-.75	.77	.80	-.88	.71
.80	.72	.79	.45	.62	.68	.21	.54	.58	.00	.50	.50	-.32	.51	.58	-.60	.62	.68	-.76	.72	.79	-.89	.54
.90	.64	.95	.68	.40	.95	.47	.18	.95	.22	.00	-	-.22	.18	.95	-.63	.41	.95	-.81	.64	.95	-.93	.05
$\rho_{yy'}$=.95																						
.50	.90	.90	.00	.88	.88	-.19	.86	.87	-.33	.85	.85	-.50	.84	.84	-.65	.84	.84	-.76	.85	.85	-.87	.90
.60	.88	.89	.12	.85	.86	-.09	.83	.84	-.24	.81	.81	-.45	.81	.81	-.63	.81	.81	-.75	.83	.83	-.87	.87
.70	.86	.88	.27	.81	.83	.04	.78	.79	-.14	.75	.75	-.39	.75	.75	-.61	.76	.77	-.75	.80	.81	-.88	.81
.80	.83	.87	.45	.75	.79	.21	.68	.71	.00	.63	.63	-.32	.63	.64	-.60	.68	.70	-.76	.75	.79	-.89	.68
.90	.79	.91	.68	.61	.82	.47	.43	.66	.22	.25	.26	-.22	.34	.51	-.63	.50	.67	-.81	.68	.86	-.93	.31

R - Raw Gain Score Reliability Coefficients

Re - Residual Gain Score Reliability Coefficients

Est - Estimated True Gain Score Reliability Coefficients

ρ_{xg} - Correlation between pretest scores and raw gain scores

* - Cannot be calculated since ρ_{xy} exceeds maximum possible value

Table 6

Raw Gain Score, Residual Gain Score, and Estimated True Gain Score Reliability Coefficients Assuming $\rho_{xx'}$=.95.

ρ_{xy}	λ=.50 R	λ=.50 Est	λ=.50 ρ_{xg}	λ=.67 R	λ=.67 Est	λ=.67 ρ_{xg}	λ=.80 R	λ=.80 Est	λ=.80 ρ_{xg}	λ=1.00 R	λ=1.00 Est	λ=1.00 ρ_{xg}	λ=1.25 R	λ=1.25 Est	λ=1.25 ρ_{xg}	λ=1.50 R	λ=1.50 Est	λ=1.50 ρ_{xg}	λ=2.00 R	λ=2.00 Est	λ=2.00 ρ_{xg}	Re
ρyy'=.75																						
.50	.65	.65	.00	.65	.65	-.19	.66	.67	-.33	.70	.72	-.50	.75	.78	-.65	.79	.82	-.76	.85	.87	-.87	.65
.60	.60	.61	.12	.58	.58	-.09	.59	.59	-.24	.63	.65	-.45	.69	.73	-.63	.75	.80	-.75	.83	.87	-.87	.58
.70	.52	.58	.27	.47	.48	.04	.46	.46	-.14	.50	.54	-.39	.60	.68	-.61	.68	.77	-.75	.80	.86	-.88	.46
.80	.42	.68	.45	.28	.44	.21	.22	.24	.00	.25	.36	-.32	.42	.68	-.60	.57	.82	-.76	.75	.90	-.89	.22
.90	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
ρyy'=.80																						
.50	.72	.72	.00	.71	.71	-.19	.72	.73	-.33	.75	.76	-.50	.79	.80	-.65	.82	.84	-.76	.87	.88	-.87	.72
.60	.67	.68	.12	.66	.66	-.09	.66	.66	-.24	.69	.70	-.45	.74	.76	-.63	.78	.81	-.75	.85	.87	-.87	.66
.70	.61	.65	.27	.56	.57	.04	.55	.55	-.14	.58	.60	-.39	.66	.70	-.61	.73	.78	-.75	.82	.86	-.88	.56
.80	.53	.67	.45	.41	.49	.21	.36	.37	.00	.38	.42	-.32	.51	.65	-.60	.63	.78	-.76	.78	.88	-.89	.36
.90	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
ρyy'=.85																						
.50	.78	.78	.00	.78	.78	-.19	.78	.78	-.33	.80	.65	-.50	.83	.83	-.65	.85	.86	-.76	.88	.89	-.87	.78
.60	.75	.76	.12	.73	.73	-.09	.73	.73	-.24	.75	.76	-.45	.79	.80	-.63	.82	.83	-.75	.87	.88	-.87	.74
.70	.70	.72	.27	.66	.67	.04	.65	.65	-.14	.67	.67	-.39	.72	.74	-.61	.77	.80	-.75	.84	.87	-.88	.66
.80	.64	.72	.45	.54	.59	.21	.49	.50	.00	.50	.51	-.32	.59	.66	-.60	.69	.77	-.76	.81	.87	-.89	.49
.90	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
ρyy'=.90																						
.50	.85	.85	.00	.84	.84	-.19	.84	.84	-.33	.85	.85	-.50	.86	.87	-.65	.88	.88	-.76	.90	.90	-.87	.85
.60	.83	.83	.12	.81	.81	-.09	.81	.81	-.24	.81	.81	-.45	.83	.84	-.63	.85	.86	-.75	.88	.89	-.87	.82
.70	.80	.81	.27	.76	.76	.04	.75	.75	-.14	.75	.75	-.39	.78	.79	-.61	.82	.83	-.75	.86	.88	-.88	.76
.80	.75	.79	.45	.68	.70	.21	.53	.64	.00	.63	.63	-.32	.68	.71	-.60	.75	.79	-.76	.83	.87	-.89	.63
.90	.68	.86	.68	.50	.73	.47	.34	.51	.22	.25	.26	-.22	.43	.66	-.63	.61	.82	-.81	.79	.91	-.93	.26
ρyy'=.95																						
.50	.92	.92	.00	.91	.91	-.19	.90	.90	-.33	.90	.90	-.50	.90	.90	-.65	.91	.91	-.76	.92	.92	-.87	.92
.60	.90	.91	.12	.89	.89	-.09	.88	.88	-.24	.88	.88	-.45	.88	.88	-.63	.89	.89	-.75	.90	.91	-.87	.89
.70	.89	.89	.27	.86	.86	.04	.84	.84	-.14	.83	.83	-.39	.84	.84	-.61	.86	.86	-.75	.89	.89	-.88	.85
.80	.86	.88	.45	.81	.82	.21	.77	.78	.00	.75	.75	-.32	.77	.78	-.60	.81	.82	-.76	.86	.88	-.89	.77
.90	.82	.88	.68	.70	.78	.47	.59	.65	.22	.50	.50	-.22	.59	.65	-.63	.70	.78	-.81	.82	.88	-.93	.52

R - Raw Gain Score Reliability Coefficients
Re - Residual Gain Score Reliability Coefficients
Est - Estimated True Gain Score Reliability Coefficients
ρ_{xg} - Correlation between pretest scores and raw gain scores
* - Cannot be calculated since ρ_{xy} exceeds maximum possible value

References

Bereiter, C. (1963). Some persisting dilemmas in the measurement of change. In C.W. Harris (Ed.), Problems in measuring change (pp. 3-20). Madison, WI: University of Wisconsin Press.

Campbell, D.T., & Stanley, J.C. (1966). Experimental and quasi-experimental designs for research. Boston: Houghton Mifflin. (Reprinted from Handbook on research in teaching, pp. 171-246, by N.W. Gage, Ed., 1963, Chicago: Rand McNally.

Corder-Bolz, C. (1978). The evaluation of change: New evidence. Educational and Psychological Measurement, 38, 959-976.

Crocker, L.M., & Algina, J. (1986). Introduction to classical and modern test theory. Fort Worth, TX: Holt, Rinehart and Winston.

Cronbach, L.J., & Furby, L. (1970). How we should measure "change"-or should we? Psychological Bulletin, 74, 68-80.

CTB/McGraw-Hill. (1987). California achievement tests forms E and F technical report. Monterey, CA: Author.

Diederich, P.B. (1956). Pitfalls in the measurement of gains in achievement. School Review, 64, 59-63.

DuBois, P.H. (1957). Multivariate correlational analysis. New York: Harper, 1957.

Engelhart, M. (1967). A note on the analysis of gains and posttest scores. Educational and Psychological Measurement, 27, 257-260.

Feldt, L.S. (1958). A comparison of the precision of three experimental designs employing a concomitant variable Psychometrika, 23(4), 335-353.

Feldt, L.S., & Brennan, R.L. (1989). Reliability. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 105-146). New York: Macmillan.

Fortune, J.C., & Hutson, B.A. (1984). Selecting models for measuring change when true experimental conditions do not exist. Journal of Educational Research, 77(4), 197-206.

Kenny, D.A. (1975). A quasi-experimental approach to assessing treatment effects in the nonequivalent control group design. Psychological Bulletin, 82(3), 345-362.

Linn, 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. (1956). The measurement of growth. Educational and Psychological Measurement, 16, 421-437.

Lord, F.M. (1958). Further problems in the measurement of growth. Educational and Psychological Measurement, 18, 437-451.

Lord, F.M. (1963). Elementary models for measuring change. In C.W. Harris (Ed.), Problems in measuring change (pp. 21-38). Madison, WI: University of Wisconsin Press.

Manning, W.H., & DuBois, P.H. (1962). Correlational methods in research on human learning. Perceptual and Motor Skills, 15, 287-321.

Martin, D.J. (1985). The measurement of growth in educational achievement. Unpublished doctoral dissertation, University of Iowa, Iowa City.

Maxwell, S., & Howard, G. (1981). Change scores-necessarily anathema?. Educational and Psychological Measurement, 41, 747-756.

Microsoft Corporation (1985-94). Microsoft Excel Version 5.0C. United States of America: Author.

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

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

Psychological Corporation (1993). Metropolitan achievement test technical manual (7th Ed). San Antonio, TX: Author.

Rachor, R.E., & Cizek, G.J. (1994, October). An empirical investigation of the reliability of gain scores. Paper presented at the annual meeting of the Mid-Western Educational Research Association: Chicago.

Richards, J.M. Jr. (1975). A simulation study of the use of change measures to compare educational programs. American Educational Research Journal, 12(3), 299-311.

Rogosa, D., Brandt, D., & Zimowski, M. (1982). A growth curve approach to the measurement of change. Psychological Bulletin, 92(3), 726-748.

Rogosa, D., & Willett, J. (1983). Demonstrating the reliability of the difference score in the measurement of change. Journal of Educational Measurement, 20(4), 335-343.

Stanley, J.C. (1967). General and specific formulas for reliability of differences. Journal of Educational Measurement, 4, 249-252.

Stanley, J.C. (1971). Reliability. In R.L. Thorndike (Ed.), Educational measurement (2nd ed., pp. 356-442). Washington, DC: American Council on Education.

Thomson, G.H. (1925). An alternate formula for the true correlation of initial values with gains. Journal of Experimental Psychology, 8, 323-324.

Thorndike, R.L. (1924). The influence of chance imperfections of measures upon the relation of initial score to gain or loss. Journal of Experimental Psychology, 7, 225-232.

Thorndike, R.L. (1966). Intellectual status and intellectual growth. Journal of Educational Psychology, 57, 121-127.

Traub, (1967). A note on the reliability of residual change scores. Journal of Educational Measurement, 4, 253-256.

Traub, R.E. (1994). Reliability for the social sciences. Thousand Oaks, CA: Sage Publications.

Veldman, D., & Brophy, J. (1974). Measuring teacher effects on pupil achievement. Journal of Educational Psychology, 66, 319-324.

Wiersma, W. (1991). Research Methods in Education. Boston: Allyn and Bacon.

Willett, J.B. (1988-89). Questions and answers in the measurement of change. In E.Z. Rothkopf (Ed.), Review of research in education 15, (pp. 345-422). Washington, DC: American Educational Research Association.

Zimmerman, D., & Williams, R. (1982a). Gain scores in research can be highly reliable. Journal of Educational Measurement, 19(2), 149-154.

Zimmerman, D., & Williams, R. (1982b). On the high predictive potential of change and growth measures. Educational and Psychological Measurement, 42, 961-968.