

APPLYING MEASUREMENT OF CHANGE

Applying the Theory on Measurement of Change
to Formative Classroom Assessment

Abstract

Current practices in classroom assessment tacitly follow a classical test theory model that assumes a summative perspective. Yet, classroom assessment is inherently formative in nature. This paper utilizes current theory on the measurement of change to outline the rudiments of a measurement theory base for formative classroom assessments. To this end, the advantages and disadvantages of two growth functions are discussed – difference scores and multi-wave data fitted to a linear function. The disadvantages of the classical conceptions of reliability when estimating individual growth are considered. While it is demonstrated that the difference score function can provide viable estimates of individual student growth, the linear, multi-wave function is preferred.

Applying the Theory on Measurement of Change to Classroom Assessment

Classroom assessment is at the heart of classroom culture. It has been shown to be the basis for grading practices (Airasian, 1994; Brookhart, 1994; Haladayna, 1999; Stiggins et al., 1989) and curriculum design (McMillan, 1997, 2000). It has even been shown to enhance student achievement (Fuchs & Fuchs, 1986). As important as classroom assessment is, it has operated without benefit of a well-articulated theory base. Perhaps more accurately, classroom assessment has operated under the tacit, albeit inappropriate auspices of classical test theory. As Taylor and Nolen (1996) note, the “paradigm” of classical test theory with its summative perspective has been inappropriately superimposed on classroom assessment. They further add that a formative paradigm is more appropriate for classroom assessment.

The distinction between formative and summative assessment was first popularized by Scriven in 1967 as part of an American Educational Research Association monograph series on evaluation. Scriven’s original point was that a distinction should be made between programs that are being formulated versus programs that have evolved to their final state. Consequently, evaluation takes on different characteristics and is interpreted differently in formative versus summative situations. This distinction was soon applied to the assessment of students. Specifically, formative assessment was defined as occurring while a trait is being learned. Summative assessment occurs at the end of a learning cycle (see McMillan, 2000). To illustrate, within the classroom context, Airasian (1994) defines formative assessments as those that “are interactive and used primarily to form or alter an ongoing process or activity. In contrast,

assessments that come at the end of a process or activity, when it is difficult to alter or rectify what has already occurred, are called summative assessments” (pp. 135, 136).

The assertion that classroom assessment is inherently formative is supported from at least two perspectives. First, given the fact that most of what goes on in a classroom is designed to form or alter student learning, summative assessment can be used only at the very end of a unit of instruction. Second, within the context of the classroom, assessments are frequently used effectively as aids to learning. In their review of some 681 publications in formative assessment, Black and Wiliam (1998) noted that frequent feedback to students via formative assessment has the potential of dramatically increasing the achievement of students. Assuming an effect size of .70 (Cohen’s *d*) for formative assessment, Black and Wiliam (1998) note that “if it could be achieved on a nationwide scale, it would be equivalent to raising the mathematics score of an average country... into the ‘top five’ after the Pacific rim countries” (p. 61). Similar findings have been reported by Wenglinsky (2000).

In short, the basic premise of this article is that classroom assessment is inherently formative in nature, yet a summative theory base has been inappropriately applied to it. Finally, this article attempts to apply the theory of the measurement of change to classroom assessment.

The Summative Bias in Current Classroom Assessment

Perhaps the best evidence that the current practice of classroom assessment is organized

around a summative perspective is the common practice of average classroom assessment scores. To illustrate, best-selling textbooks on classroom assessment either implicitly or explicitly communicate the message that summary scores for students at the end of a grading period should be computed using weighted or unweighted averages of classroom tests. For example, Airasian (1994) provides explicit examples of how teachers should average test and homework scores obtained throughout a unit of instruction to compute summary scores for students. These summary scores are then translated into final grades (p. 316). Haladyna (1999) provides similar advice. Averaging assessment gathered throughout a unit of instruction to obtain summary scores for students appears to be standard operating procedure in the classroom.

Averaging assessment scores makes perfect sense from a summative perspective. To illustrate, averaging scores on a set of parallel assessments provides an unbiased estimate of an individual's true score. Indeed, the construct of true score is commonly defined in terms of the expected value of parallel tests. Lord (1959) explains that a true score is "frequently defined as the average of the scores that the examinee would make on all possible parallel tests if he did not change during the testing process" (p. 473). Magnusson (1966) describes true score in the following way: ". . . the true score which can be predicted with complete certainty from the latent continuum is the same for every individual from one parallel test to the other" (p. 63). Gulliksen (1950) defines true score for a given student as ". . .the limit that the average of his scores on a number of tests approaches, as the number of parallel tests. . . increases without limit" (p. 28).

As the quotations above indicates, the driving force behind averaging as a way of

estimating an individual's true score is the assumption that the true score for a given individual on a given trait is constant from assessment to assessment. Indeed, the construct of identical true scores is at the heart of definitions of parallel tests (see Lord & Novick, 1968). From a summative perspective, then, it follows that averaging over multiple assessments will decrease or theoretically remove the error score component leaving one with a viable estimate of an examinee's true score. As Magnusson (1966) notes: "The greater number of parallel tests we administer, the greater the chances are that the random errors will cancel each other out. The sum of the error scores will be zero for an infinite number of parallel tests" (p. 64).

Modeling True Score Change

The problem with averaging within a formative model is that by definition, an individual's true score changes from assessment occasion to assessment occasion. That is, as students learn, their true score increases. If one is to use the formative data points available to a teacher throughout a unit of instruction, then, a theory base must be used that assumes a monotonically increasing true score for each student. Fortunately, such a theory base has been articulated within the literature on the measurement of change. Willett (1988) has documented the rich history of this literature, but to date, it has not been applied to classroom assessment.

As described by Willett (1985), the basic measurement model for this endeavor is

$$X_{ip} = F_p(t_i) + e_{ip} \quad (1)$$

where the subscript i denotes the occasion of measurement, t_i is the time at which the i^{th} occasion of measurement occurred, and the subscript p indicates the person being measured. The symbol F_p represents true status for person p , and the parenthetical inclusion of the time at which the i^{th} measurement occurred indicates that F_p is changing (growing) over time. Consequently, the notation $F_p(t_i)$ represents a function describing the true status of individual p at varying times. e_{ip} represents the measurement error associated with person p at time i .

Contrast this formative measurement model with the summative model from classical test theory:

$$X_p = F_p + e_p \quad (2)$$

Here F_p is the true status of person p . The absence of the parenthetical expression (t_i) illustrates that the classical measurement model assumes a fixed true status.¹ From equation 1 above, one might infer that a critical aspect of constructing a viable formative measurement model is to identify the most appropriate growth function, $F_p(t_i)$. Below, two growth functions are considered: the difference score function and the multi-wave linear function.

The Difference Score Function

The simplest growth function that might be used as the basis for classroom assessment is

the difference between initial and final status. In such a situation, a classroom teacher would utilize an initial test of student achievement on the content to be covered in a unit of instruction. Each student's score on this initial assessment would be subtracted from the student's score on an end-of-unit test under the assumption that initial and final tests are parallel. The research on the sensitization effects of a pretest indicates that classroom teachers might even use the same test for both occasions (Wilson & Putnam, 1982). The syntactic (i.e., mathematical) model for this growth function is

$$D_p = X_{fp} - X_{1p} \quad (3)$$

where X_{fp} is the final measurement for person p (e.g., the final assessment during a unit of instruction) and X_{1p} is the first measurement for person p . There is a simple relationship between the observed difference score and the underlying change in true score that has occurred between t_f and t_1 :

$$D_p = \Delta_p + e_p^* \quad (4)$$

Where $\Delta_p = F_p(t_f) - F_p(t_1)$ and $e_p^* = e_{fp} - e_{1p}$. As a consequence of the assumption of independence of error terms, one can conclude the e^* is normally distributed with zero mean and variance $2\sigma_e$. Additionally, it can be shown that the observed difference score for individual p is an unbiased estimate of the quantity Δ_p . Rogosa et al. (1982) emphasize the fact that the observed difference score is an unbiased estimate of the true difference score regardless of the magnitude

of the measurement error. This is necessarily so because the expected value of the observed difference score for person p is the true difference score for person p . Willett (1988) notes that in spite of “this obvious and optional statistical property. . .”, the difference score has been criticized “so thoroughly and continuously over the years that investigators have become wary of its use. . .” (p. 366) most notably for its perceived low reliability. While it is true that difference scores will frequently exhibit low reliability, this does not necessarily mean they exhibit poor precision at the level of individual difference scores. An examination of the conceptual underpinnings of the reliability of difference scores provides some insight into the issue.

The population reliability of difference scores $\rho(D)$ is defined as the ratio of variance of Δ_p to D_p over all individuals in the population. Given that $\rho(D)$ is expressed in terms of observed differences, Willett offers the following formula which requires no simplifying assumptions

$$\rho(D) = \frac{\mathbf{s}_{x_1}^2 p(x_1 x_1) + \mathbf{s}_{x_f}^2 p(x_f x_f) - 2\mathbf{s}_{x_1} \mathbf{s}_{x_f} p(x_1 x_f)}{(\mathbf{s}_{x_1}^2) + (\mathbf{s}_{x_f}^2) - 2\mathbf{s}_{x_1 x_f} p(x_1 x_f)} \quad (5)$$

Here $p(x_1 x_1)$ and $p(x_f x_f)$ are the population reliabilities for x_1 and x_f , $p(x_1 x_f)$ is the population correlation between initial scores (x_1) and final status scores (x_f). The fact that the term containing the correlation between initial and final status is subtracted in both numerator and denominator is the reason commonly given for assumed poor reliability of difference scores. As Willett explains,

Because the variances in the first two terms of the numerator are multiplied by their

respective reliabilities and are, therefore, smaller than the equivalent terms in the denominator, the subtraction of the term containing the between-wave correlation in both the numerator and the denominator ensures that when $\rho(x_i x_f)$ is large and positive, then $\rho(D)$ will be low. And, as the correlation of initial and final status is frequently misinterpreted as an index of construct validity, authors are apt to report that the difference score cannot be both reliable and valid simultaneously (p. 369).

To illustrate the impact of a high correlation between initial and final status on the reliability of difference scores, consider the situation in which $\rho(x_i x_f)$ is zero. The term subtracted in numerator and denominator becomes zero, rendering equation 5 to be

$$\rho(D) = \frac{\mathbf{s}_{x_i}^2 \rho(x_i x_i) + \mathbf{s}_{x_f}^2 \rho(x_f x_f)}{\mathbf{s}_{x_i}^2 + \mathbf{s}_{x_f}^2} \quad (6)$$

In equation 6, the reliability of the difference scores reduces to the variance-weighted average of $\rho(x_i x_i)$ and $\rho(x_f x_f)$. If we assume that $\sigma^2(x_i) = 4$, $\sigma^2(x_f) = 9$, $\rho(x_i x_i) = .90$, and $\rho(x_f x_f) = .80$, then, $\rho(D) = .83$. However, if we assume a moderate correlation between initial and final status of .50, then $\rho(D)$ is reduced to .69.

Given that the correlation between tests of initial and final status are usually quite high, the practical implication is that the computed reliability of difference scores will almost always be low. As Feldt and Brennan (1989) note:

When one looks at the reported data for standardized achievement tests, as in

reading, it is not uncommon to find [difference score] reliabilities in the neighborhood of .33 or lower. Yet, no one would deny that the typical fifth-grade student makes considerable progress in a year's time. (p. 119)

Willett provides further insight into the reliability of difference scores using the perspective of the homogeneity (or lack thereof) of individual growth rates. He notes that a more illustrative formulation of the reliability of difference scores is

$$p(D) = \frac{\mathbf{s}_{\Delta}^2}{\mathbf{s}_{\Delta}^2 + 2\mathbf{s}_e^2} \quad (7)$$

Here it is clear that the reliability of the difference score increases as the differences in true change increases among individuals. As Willett notes,

Thus, the greater the individual differences in true growth, the greater the reliability of the difference score. Where there are no individual differences in true growth to detect (i.e., when σ_{Δ}^2 is equal to zero and all the individual true score growth trajectories are parallel), the reliability of the difference score can only be zero, regardless of the precision with which measurement has been carried out. (p. 320)

Ultimately, an analysis of the theoretical underpinnings of the reliability of difference scores leads one to the conclusion that this construct is not well-suited to the task of examining the

viability of individual difference scores. By definition, "Reliability is a measure of inter- individual differentiations and can only be defined over a group or population" (Rogosa et al., 1982, p. 730).

Given that heterogeneity of growth rates is a prerequisite for high reliability of difference scores, assessment situations in which growth rates are homogeneous will produce low reliabilities, but tell one little about the precision of measurement. As Rogosa et al. note:

Although individual differences in growth are necessary for high reliability, the absence of such differences does not preclude meaningful assessment of individual change. (p. 731)

In summary, despite the characteristically low reliabilities associated with them, initial to final status difference scores are a viable candidate for the growth function in a formative measurement model. Specifically, teachers might administer a comprehensive assessment at the beginning and end of a unit, and use the difference scores as viable estimates of student academic growth. This said, a multi-point or multi-wave approach has a number of advantages over a difference score approach.

The Multi-Wave Linear Function

While the above discussion supports the use of difference scores as estimates of students' learning, the practice is still problematic. Even though estimates of the reliability of difference

scores are not well suited to the assessment of individual change, they should still be estimated.

However, such estimates require information not easily obtained by the classroom teacher. As

Willett (1988) notes:

In addition to the sample variances and correlation provided by the two waves of growth data, the investigator must also have two supplementary pieces of information: the estimated reliability of observed status on each of the two occasions of measurement. Either this supplementary information must be obtained externally to the growth investigations (i.e., from a test manual or a previous empirical reliability study), or duplicate measurements of observed status must be made on each subject at each of the two time points to permit in situ estimation of required reliability. (p. 371)

Obviously, these requirements for supplemental information do not fit well in the context of classroom assessments. An approach that more readily allows for the estimation of a reliability index that is more meaningful to the estimation of individual growth is the use of a multi-point or multi-wave data. Additionally, multi-wave data provide for a better estimate of true growth than do difference scores:

. . .taking a snapshot of individual status on each of two occasions does not permit the investigator to visualize the intricacies of the underlying individual growth with any great certainty. . .Indeed, to measure individual growth adequately, more

information on that growth in the form of multi-wave data is required. When multi-wave data are available on each of the subjects in the sample, the investigator can examine detailed empirical growth – trajectory plots that summarize the observed growth of each individual over time. (Willett, 1988, pp. 384-385)

In a classroom situation, multi-wave data might be collected by a teacher administering multiple parallel comprehensive tests throughout a unit of instruction. For example, a teacher might construct two parallel comprehensive tests and then alternate their use. Two administrations of each test would provide four data points for each student.

One way to conceptualize the advantages of multi-wave data is to think of the difference score as an estimate of the regression weight for the linear function representing an individual's growth when two data points only are available. Indeed, when only two data points are available, the regression weight for the straight-line function for an individual student is

$$\frac{x_f - x_l}{t_f - t_l} \quad (8)$$

This is directly proportioned to the raw difference score (see Willett, 1988, p. 385; Rogosa et al., 1982, p. 728).

Modeling growth using multi-wave data is based on the assumption that learning over time follows a definable mathematical function and that observed measurements over time are

imprecise estimates of the growth function.

When an individual is growing, it is as though the underlying true growth is continuing smoothly and unobserved overtime, but periodically, the investigator observes the growth with some fallible measuring instrument. In this way, the individual's observed growth record is assembled and it consists of a chronological series of discrete measurements each of which is an unknown combination of true status and measurement error. What is of fundamental interest to the investigator, of course, is the underlying, continuous true growth trajectory; the multiple entries in the observed growth record are simply a fallible lens through which the true individual growth is viewed. (Willett, 1988, pp. 387-388)

Mead and Pike (1975) list a variety of algebraic functions that might be used to model the true growth function for individuals as do Rogosa et al. (1982). Among others, these include a straight-line function, a quadratic function, and a negative exponential function. The most intuitively obvious model – the straight-line function – will be the focus of the discussion below, although other functions can certainly be applied to classroom assessment.

A straight-line function has many advantages, not the least of which is the straightforward nature of the computations necessary to estimate parameters. Also, a case can be made that even if the true function describing an individual person's growth trajectory is nonlinear, the linear function provides a viable perspective. For example, Seigel (1975) has demonstrated that even

when a second degree polynomial is the appropriate function, the straight-line fit is a good estimate of the average rate of change.

To illustrate the use of a straight-line function with multi-wave classroom data, consider

Table 1:

Table 1 Here

These data were collected by a classroom teacher over a nine-week period of time. Two comprehensive parallel tests were constructed and alternately administered at intervals that were approximately equal, with the first administration occurring on the first day of class and the final administration on the last day of class.

In the context of the present discussion, the straight-line function modeling these data would be represented as

$$F_p(t) = F_p(t^*) + B_p(t - t^*) \quad (9)$$

where $F_p(t^*)$ represents true initial status, $(t - t^*)$ represents the time differential between occasion of measurement t_i and occasion t^* , and B is the regression coefficient which represents the (constant) rate of change from occasion to occasion.ⁱⁱ

Given the use of the linear function, the basic measurement model now becomes

$$X_{ip} = F_p(t^*) + B_p(t - t^*) + e_{ip} \quad (10)$$

For the purposes of estimating individual growth, the key parameter is B_p – the coefficient for the regression of the observed score on the change in occasion of measurement, or the rate of change in the observed score for one unit of change in occasion of measurement (assuming equal distances between units). The precision of this estimate is analogous to the precision of estimate of an individual's true status at a given point in time.

Table 2 depicts the linear equations for the data reported in Table 1:

Table 2 Here

In all, there are 21 equations – one for each student. For each equation, Table 2 depicts the estimated slope for each equation along with the standard error of the slope, the coefficient of determination, and the residual variance. Given that these data represent scores for students on tests given over a nine-week period of time, the predicted score for the final occasion of measurement is a viable estimate of the student's true score at the end of the nine weeks. This is depicted in column 5 of Table 2.

The predicted final score for an individual (as opposed to the observed final score) has the

advantage of utilizing the multiple data points inherent in formative assessment. The multiple data points also allow for estimations of precision and reliability. Willett (1988) explains that the relationship between the population variance of growth rates (s_B^2) and the sample variance ($s_{\hat{B}}^2$) is

$$s_{\hat{B}}^2 = s_B^2 + \frac{s_e^2}{SST} \quad (11)$$

where SST is the sum of squares for the observation times – the squared deviations of the observation times about their mean ($S[t_i - \bar{t}]^2$). This is a measure of the spread of occasion of measurement. s_e^2 is the variance due to measurement error. Equation 11 illustrates that the sample variance of the growth rates will overestimate the true variance of the growth rates.

Given that multi-wave data are available, measurement error variance for each person can be computed and, consequently, used in the estimate of the population variance of growth rates. Specifically, given that the growth model fitted to the data is a correct one, the differences between the observed scores and predicted scores estimate the measurement error on each occasion of measurement. As Willett notes,

The magnitude of the measurement error can be found directly from these residuals.

Under the assumptions of this paper, an estimate of s_e^2 can be found quite simply by summing the squared residuals. . . across occasions and persons, and then

dividing by the total degrees of freedom. (p. 403)

By algebraic manipulation, the sum of the squared residuals can be shown to be equal to the simple average of the *MSE* for each person. Therefore,

$$\hat{s}_e^2 = \frac{\sum_{p=1}^n MSE_p}{n} \quad (12)$$

Using the data in Table 2, we compute \hat{s}_e to be 17.97. The sample variance of the slopes is 7.88 and $SST = 5$. Using these quantities, an estimate of the population variance of growth rates is

$$\hat{s}_B^2 = s_B^2 - \frac{\hat{s}_e^2}{SST} \quad (13)$$

Therefore, the estimate of the population variance of growth rates is 4.29.

Estimating the Reliability of Growth Rates

From a semantic perspective, the reliability of growth rates is defined as the ratio of variance of the true growth rates over the variance of the observed growth rates. Since the variance of the observed growth rates can be computed as well as a viable estimate of the variance of the true growth rates, the population reliability of the growth rates can be estimated via the following formula:

$$p(\hat{B}) = \frac{\mathbf{s}_B^2}{\mathbf{s}_B^2 + \frac{\mathbf{s}_e^2}{SST}} \quad (14)$$

Using the estimates from equations 12 and 13, this reliability is .591 for the data in Table 2. As is the case with reliability estimates for difference scores, one must be cautious in the interpretation of the reliability of growth rates. Willett points out that all else being equal, the larger the true variance in growth rates, the larger will be the reliability.

Thus, in a population where there is plenty of criss-crossing of true growth trajectories, considerable reliability is possible in practice. On the other hand, if there is no inter-individual differences in the rate of true growth ($\mathbf{s}_B^2 = 0$), then all of the true growth trajectories will be parallel and the growth reliability can only be zero, regardless of the precision with which measurement has been achieved. (Willett, 1988, p. 404)

In summary, the use of formative assessments obtained over time in the classroom allows not only for estimation of individual students' true scores at the end of a learning period, but it also allows for the estimation of individual growth trajectories, estimation of measurement error for individual students, and estimation of the reliabilities of the growth trajectory for the class considered as a group.

Conclusions

This article began with the basic assertion that the appropriate model for classroom assessment is a formative one, yet, the common practice of averaging to obtain summary scores for individual students is inherently summative in nature. Fortunately, well-established models for the measurement of change have been articulated but as yet not applied to the classroom context. Two models were addressed here – the difference score model and the multi-wave linear model.

Classroom teachers could apply these models by simply designing and administering a pretest, posttest, and one or more intervening tests (if the multi-wave linear model is used). Certainly, classroom teachers would not be expected to compute the quantities associated with the difference score function or the linear multi-wave function. However, computerized grade books currently exist that perform such calculations requiring that teachers only enter formative test scores for students.

Finally, Rogosa et al. (1982) provide a number of “mottos” for the measurement of individual change that apply nicely to the present discussion regarding classroom assessment.

- Two waves of data are better than one but maybe not much better. Two data points provide meager information on individual change and, thus, the measurement of change often will require more than the traditional pre-post data.

- When only two waves of data are available, the difference score is a natural and useful estimate of individual change.
- There is more than one way to judge a measure of change. Reliability is not the “be all and end all” in the measurement of change. Statistical properties are important.
- Low reliability does not necessarily mean lack of precision.
- The difference between two fallible measures can be nearly as reliable as the measures themselves.

Using the well-articulated theory of measurement of change, these mottos can be executed to form the basis of a formative model of classroom assessment.

References

Airasian, P. W. (1994). *Classroom assessment* (2nd ed.). New York: McGraw Hill.

Black, P., & William, D. (1998). Assessment and classroom learning. *Assessment in Education*, 5(1), 7-75.

Brookhart, S. M. (1994). Teachers' grading: Practices and theory. *Applied Measurement in Education*, 7(4), 279-301.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Lin (Ed.), *Educational Measurement* (3rd ed., pp. 105-146). New York: Macmillan Publishing Company.

Fuchs, L. S. & Fuchs, D. (1986). Effects of systematic formative evaluation: A meta-analysis. *Exceptional Children*, 53, 199-208.

Gulliksen, H. (1950). *Theory of mental tests*. New York: John Wiley & Sons.

Haladyna, T. M. (1999). *A complete guide to student grading*. Boston, MA: Allyn & Bacon.

Lord, F. M. (1959). Problems in mental test theory arising from errors of measurement. *Journal of the American Statistical Association*, 472-479.

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

Magnusson, D. (1966). *Test theory*. Reading, MA: Addison & Wesley.

McMillan, J. H. (1997). *Classroom assessment: Principles and practice for effective instruction*. Boston: Allyn & Bacon.

McMillan, J. H. (2000). *Basic assessment concepts for teachers and administrators*. Thousand Oaks, CA: Corwin Press.

Mead, R., & Pike, D. J. (1975). A review of response surface methodology from a biometric viewpoint. *Biometrics*, 31, 803-851.

Rogosa, D. R., Brandt, D., & Zimowsky, M. (1982). A growth curve approach to the measurement of change. *Psychological Bulletin*, 90, 726-748.

Scriven, M. (1967). The methodology of evaluation. In R. F. Stake (Ed.), *Curriculum evaluation: American Education Research Association Monograph Series on Evaluation, No. 1*, (39-83). Chicago: Rand McNally.

Seigel, D. G. (1975). Several approaches for measuring average rate of change for a

second degree polynomial. *The American Statistician*, 29, 36-37.

Stiggins, R. J., Frisbie, D. A., & Griswold, P. A. (1989). Inside high school grading practices: Building a research agenda. *Educational Measurement: Issues and Practices*, 8(2), 5-14.

Taylor, C. S., & Nolen, S. B. (1996, November 11). What does the psychometrician's classroom look like? Reframing assessment concepts in the context of learning. *Education Policy Analysis Archives*, 4(17), 1-39. Retrieved September 1 from <http://olam.ed.asu.edu/epaa/v4n17.html>

Wenglinsky, H. (2000). *Teaching matters: Bringing the classroom back into discussions of teacher quality*. Princeton, NJ: Educational Testing Service.

Willett, J. B. (1985). *Investigating systematic individual difference in academic growth*. Unpublished doctoral dissertation, Stanford University, Palo Alto, CA.

Willett, J. B. (1988). Questions and answers in the measurement of change. *Review of Research in Education* (Vol. 15, pp. 345-422). Washington, DC: American Educational Research Association.

Wilson, V. L., & Putnam, R. R. (1982). A meta-analysis of pretest sensitization effects on experimental design. *American Educational Research Journal*, 19(2), 249-258.

Endnotes

i) According to Lord and Novick (1968, pp. 36 and 38), the classical test theory measurement model is defined by a set of theorems that address the expected values and correlations of X , F , and e , for samples of people across occasions of parallel measurements. For this discussion, the major point of interest is the contrast between classical test theory's use of a single true score for a given person within the measurement model, and change theory's use of multiple true scores [i.e., a true score growth function] in the measurement model.

ii) Willett (1988, p. 390) notes that a more thoughtful approach is to redefine t^* as the average value \bar{t} of the T times of observation. This increases the precision of the intercept estimate because the estimate is taken in "midstream" of the observations. The subsequent measurement model would be

$$x_{ip} = F_p(\bar{t}) + \mathbf{s}_p(t - \bar{t}) + e_{ip}$$

Table 1

Multi-word Classroom Data

Student ID	Score Waves			
	1	2	3	4
1	65	77	75	82
2	75	74	82	91
3	82	88	86	91
4	61	82	79	83
5	56	61	82	83
6	51	76	80	82
7	85	90	96	97
8	72	77	82	81
9	76	77	82	71
10	52	57	73	77
11	86	87	92	95
12	65	75	77	75
13	71	85	90	95
14	62	67	71	72
15	42	51	50	61
16	65	70	72	75
17	66	81	79	85
18	52	65	71	71
19	61	66	69	74
20	54	60	61	62
21	71	75	68	74
<i>M</i>	65.24	73.38	77.00	79.86
<i>SD</i>	11.78	10.66	10.49	10.18

Table 2

Parameter Estimates for Multi-wave Data

Student ID	Estimated Slope B	Standard Error of Slope $SE B$	Coefficient of Determination R^2 %	Predicted Wave 4	Residual Variance MSE
1	4.90	1.81	78.60	82.10	16.35
2	5.60	1.68	84.80	88.90	14.10
3	2.50	1.07	73.10	90.50	5.75
4	6.30	3.47	62.30	85.70	60.15
5	10.20	2.62	88.31	85.80	34.40
6	9.70	3.88	75.80	86.80	75.15
7	4.20	.76	93.80	98.30	2.90
8	3.20	1.04	82.60	82.80	5.40
9	-1.00	2.34	.08	75.00	28.00
10	8.80	1.73	92.90	77.70	14.90
11	3.20	.53	94.80	94.80	1.40
12	3.20	1.92	58.20	77.80	18.40
13	6.80	1.94	85.90	94.70	18.90
14	3.40	.65	93.20	73.10	2.10
15	5.00	1.59	86.20	59.40	12.60
16	3.20	.42	96.60	75.30	.90
17	5.50	2.23	74.60	86.00	25.75
18	6.30	2.06	82.40	74.20	21.15
19	4.20	.28	99.10	73.80	.40
20	2.50	.87	80.60	63.00	3.75
21	.20	1.73	.01	72.30	14.90