DOCUMENT RESUME

ED 379 310 TM 022 671

AUTHOR Stone, Gregory Ethan
TITLE The Historical Development of Fit and Its Assessment in the Computer Adaptive Testing Environment.
PUB DATE Oct 94
NOTE 25p.; Paper presented at the Annual Meeting of the Midwestern Educational Research Association (Chicago, IL, October 1994).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Adaptive Testing; *Computer Assisted Testing; Educational Assessment; Educational History; *Goodness of Fit; Models; Regression (Statistics); *Responses; Simulation; Statistical Studies; *Test Construction; Test Format
IDENTIFIERS *Rasch Model

ABSTRACT

The quality of fit between the data and the measurement model is fundamental to any discussion of results. Fit has been the subject of inquiry since as early as the 1920s. Most early explorations concentrated on assessing global fit or subset fits on fixed length, traditional paper and pencil tests given as a single unit. The detection of aberrant response patterns in the new computer adaptive format is vital to the continued establishment of confident, quality measures. It would appear that detection strategies emphasizing the effectiveness of the targeting process are more important indicators of aberration during review than are traditional fit statistics. In fact, the targeting issue becomes a fit issue. For the investigations described, three simulated examinee records were selected, and each simulated examinee took a computer adaptive test of previously Rasch calibrated items. This exploratory study suggests that both regression and standardized measure change approaches may be viable techniques for the detection of response alteration patterns that are questionable. The educational importance of misfit is discussed. Three tables and five figures present simulation results. (Contains 18 references.) (SLD)

TM

The Historical Development of Fit and Its Assessment in the Computer Adaptive Testing Environment

Gregory Ethan Stone
The National Certification Corporation for the Obstetric, Gynecological and Neonatal Nursing Specialties

Paper presented at the Midwest Educational Research Association Annual Meeting, Chicago, Illinois, October 1994

Perspectives

The quality of fit between the data and the measurement model is fundamental to any discussion of results. Our primary goal in administering a test is to arrive at some clear understanding of an individual's knowledge about the latent-variable defined by the test items. We begin by supposing that the items included on the exam are adequate to cover the variable defined. This latent-variable is not however a concretely measurable object but rather a hypothetical construct which we specify to be represented by the limited number of items on the exam. Given the indirectness of the measurements, we must carefully assess the approach taken to each item by the examinee to determine whether or not our instrument and our subject are interacting to produce credible results.

As early as the 1920s, Thurstone recognized a need to identify and exclude from analysis individual behavior patterns that were so inconsistent with the model in use as to make reasonable measurement impossible.

> ...one must expect that some subjects will do their task in a perfunctory or careless manner....It has seemed desirable, therefore, to set up some criterion by which we could identify those individual records which were so inconsistent that they should be eliminated from our tabulations. (Thurstone, 1929)

Response patterns which violate the specifications of the measurement model in use, produce results which are highly suspect. There are at least two minimal requirements that define a good test "fit" between test and person. First, more able examinees should have higher probabilities of success on items than less able

examinees. Second, easier items should be answered correctly more often than harder items. (Mead, 1976).

These considerations arise from the probabilistic representation of the original deterministic Guttman (1944) scalogram pattern. The observed data matrix presented in Figure 1 approximates the Guttman pattern expected. The circled responses define those which are unexpected. Those to the lower right represent unexpectedly correct answers on difficult items from examinees of relatively low ability. Conversely, those to the upper left represent unexpectedly incorrect responses to easier items from examinees of higher abilities.

Cronbach (1946) sought to qualitatively describe a variety of patterns which did not follow the probabilistically determined format expected. He described six unique and observable behavior patterns, two of which were most important to high stakes multiple choice testing circumstances: Guessing and Carelessness. Guessing, he reckoned, was perhaps a result of the notion that it is better to select any answer (to guess) than to leave a response blank. Cronbach suspected that this guessing was worthwhile if all examinees were encouraged to follow the practice. Carelessness he surmised often resulted from the limited time frame in which examinees are generally expected to complete the test (Smith, 1982).

Wright and Panchapakesan's (1969) global fit statistic based upon the magnitude of departure from what is expected in the probabilistic model was the first explicit and objectively useful method for assessing inconsistency in behavior

patterns. It has served as the basis for the construction of a variety of more specific person and item fit statistics. The Wright and Panchapakesan statistic (1) and corresponding variance (2) are based upon the requirement of the Rasch model that all score groups demonstrate statistically equivalent item difficulty estimates.

$$E(\Sigma X_{vi}) = \Sigma P_{vi} \cong n_r P_{ri} = n_r \left[\frac{\exp(b_r - d_i)}{1+\exp(b_r - d_i)}\right] \quad (1)$$

$$Var(\Sigma X_{vi}) = \Sigma P_{vi}(1-P_{vi}) = n_r P_{ri}(1-P_{ri}) \quad (2)$$

P = probability of a correct response
b = person ability
d = item difficulty

Mead (1976) continued in a similar direction identifying what he considered to be three observable conditions under which person responses may not appropriately fit the model. The conditions he set forth were: (1) guessing, (2) practice and (3) test bias. An examination of the standardized residual distribution (Z-scores) derived from:

$$Z_i = \frac{X_i - P_i}{\sqrt{P_i(1-P_i)}}$$

allowed for easy detection of these aberrant patterns.

Implicit in both the relationships is the requirement that the probability of a correct response be a function of person ability and item difficulty only. Based upon this understanding of P_{vi} (probability for person *v* on item *i*) established in

the independent, logistic relationship demonstrated by the Rasch model, Mead (1979) considered three different types of person fit assessments from the analysis of residual data: total fit, within groups fit and between groups fit.

Mead's **Total fit** statistic was derived from the sum of squared residuals across the entire test:

$$\Sigma Z_i^2 = \Sigma \left[\frac{(X_i - P_i)^2}{P_i(1-p_i)} \right]$$

where X_i is the sum of all person-item interactions (x_{vi}) for each score. The result Mead suggested should be treated like a Chi-square (X^2) statistic with L-1 degrees of freedom. Wright (1980), concerned with the effects of extreme outliers on the total fit, reformulated the measure by using a ratio formed from two separate and independent estimates of the variance of the residual.

$$\Sigma Z_i^2 = \frac{\Sigma (X_i - P_i)^2}{\Sigma P_i(1-p_i)}$$

This formulation produced results comparable to those of the original estimate, but because each summation was independent, it was less sensitive to outliers that increase greatly during the summation and squaring procedure.

Mead's within group and between group fit statistics were comparable to the total fit but examined the squared residuals for selected groups or subsets of items rather than the total test. These subset fit statistics allowed for estimation of the person fit within and across given item subsets (as measures of fit invariance)

respectively. The notion of fit remained largely explained in these terms until the advent of computer adaptive testing.

Computer Adaptive Testing

Most investigations of fit up to this point, have concentrated on assessing global fit or subset fits on fixed length, traditional paper and pencil tests given as a single unit. These principles have relied upon the aberrant response patterns observable in a Guttman scalogram covering the entire test.

In 1960, Rasch concluded that if items of the test were mutually conformal, it would be possible to assess person ability by means of a test that was composed of items with similar difficulties. He continued by remarking that "it would...be necessary to have several tests of varying difficulty available in order that every person could be tested by a test of reasonable degree of difficulty for him". (Rasch, 1960) The chaining described is in theory very similar to the targeting strategy employed by computer adaptive testing.

The foundational computer adaptive testing algorithm is based upon the probabilistic model, which directs questions targeted to person ability, allowing for some fixed percent of probability of a correct response. By design if the relationship between item difficulty and person ability is thusly so controlled, any globally derived fit statistic of the sort previously discussed will be of little help in assessing fit. We would expect that the value of the statistic as here defined

would be controlled out of usefulness.

If computer adaptive testing succeeds in its targeting goal, and there is every indication that it does, then perhaps the question of misfit has become irrelevant altogether. Since one or two unexpectedly correct or incorrect answers throughout the test only negligibly effect total fit and in the adaptive testing environment, the computer algorithm immediately adjusts item selection to compensate for the new interim ability, perhaps the question has become irrelevant.

But when the simple computer adaptive person-item interaction is changed by adding other dimensions the question of fit once again arises. The dimension of review, allowing examinees to go back and alter responses at the end of the test, poses new questions and problems. Critics of computer adaptive testing have suggested that by allowing examinees the opportunity to review and alter responses, the adaptive test essentially becomes prone to "test strategy cheating". That is, by understanding the item selection algorithm, examinees could conceivably outwit the examination. The evidence regarding review clearly indicates that in practical situations this is not the case (Stone and Lunz, 1994; Bergstrom and Lunz, 1991; Lunz, Bergstrom and Wright, 1991). Instead, it has been shown that allowing review is not a psychometric liability, and that in fact, allowing review may produce a better estimate of examinee ability without the confounding of unintentional keystroke errors, initial examinee anxiety and other non-test related difficulties.

The detection of aberrant response patterns in the new computer adaptive format which allows for review is vital to the continued establishment of confident and quality measures. In general, it would appear that detection strategies emphasizing the effectiveness of the targeting process are more important indicators of aberrations during review, than are traditional fit statistics. In fact, the targeting issue *becomes* a fit issue by asking "is the examinee reviewing and altering responses in an expected manner." That is, reviewing and making alterations without causing major person ability-item difficulty mis-targeting.

Exploration

For these initial investigations, three simulated examinee records were selected from a pool of over 150. Each simulated examinee took a computer adaptive test, with a fixed length of 50 items. The test was comprised of previously Rasch calibrated items.

The CAT ADMINISTRATOR (Gershon, 1990) used the PROX estimation method (Wright & Stone, 1979) in the item selection algorithm.

The Rasch (1960/1980) model (Wright, 1977) was used to calibrate items and estimate examinee measures. Sample sizes were not large enough to meet Lord's (1983) requirements for two- and three- parameter models. There is evidence that the Rasch model yields more reliable examinee measures for small sample sizes (Green, Bock, Humphreys, Linn & Reckase, 1984). Examinee measures estimated with the Rasch model and the two- and three-parameter

models correlate above .90 when tests are administered under a computer adaptive algorithm (Olsen, Maynes, Slawson and Ho, 1986).

Upon completion of the test, the simulated examinee was allowed to review and alter any or all responses. Three simulations were extracted for this investigation: (1) an examinee making no alterations; (2) an examinee making 8 alterations (2 from correct to incorrect and 6 from incorrect to correct); and (3) an examinee making 9 alterations (all from incorrect to correct). There is some evidence that examinees tend to review in a manner resembling the second pattern on high-stakes certification examinations (Stone & Lunz, 1994). The simulations chosen represent a general spectrum from no alterations (not problematic) to 9 alterations in a consistent upward trend (possibly problematic).

Global fit was assessed using the common within and between fit statistics. Table 1 presents the results for each examinee.

Insert Table 1 about here

As expected, there were no detectable problems with the Infit (mean square) or the Outfit (mean square) for any of the three examinees. All were well within the acceptable and expected range.

When thinking about the relationship between before and after review measures we are inevitably led in some fashion or another to consider targeting.

One possible approach to illuminate the success of targeting after review uses a linear regression technique. It is clear that there is a perfect relationship between before and after review measures for individuals who do not make any response alterations. This being the case, response alterations in one particular direction (as would be the case in a cheating strategy) should be detectable by a simple regression line, based on the relationship between the observed and expected post-review responses.

Insert Table 2 and Figures 2-5 about here

Table 2 and Figures 2 through 5 explore this approach. When a typical review pattern is encountered (as seen in examinee 2 and described by Stone and Lunz, 1994) a strong and significant linear relationship between predicted and actual measures post-review appears to be maintained. If the relationship had changed in some fashion, we would expect the slope of the line to change. This was not the case with examinee 2.

Examinee 3 however, was very different. This examinee altered 9 responses from incorrect to correct. This amounts to an 18% shift in the examinee response pattern. It is apparent that the linear relationship is not maintained. Further, when inspecting the probability plot of Figure 5 there is an observable directional shift in the response pattern.

The regression method is apparently somewhat successful at detecting a response alteration pattern that is unidirectional, as would be the case in a cheating situation, and it appears to be a fairly sensitive predictor. In the case of examinee 3, only 18% of the items were altered throughout the exam, yet the loss of the linearity was very evident. However, although 16% of the responses in examinee 2 were altered, the pattern of alteration did not destroy the linear relationship. This predictive ability (maintaining the relationship with "expected" patterns and changing it with "unexpected" patterns) appears to be a positive aspect of the regression technique.

Another approach to the detection quandary examines the standardized changes in measure directly. The equation:

$$\sum \frac{(B_j - B_{j-c})^2}{SE_j^2 - SE_{j-c}^2}$$

where B = ability of examinee, j = an item in sequence, and c = some number of items, standardizes the changes of measure at certain specified intervals. To test the operation of this method, intervals were established by tens, at item 10, 20, 30, 40 and 50. Table 3 presents the results obtained.

Insert Table 3 about here

As with the regression method, the standardized measure change equation

easily detects the response alterations. Since they succeed in producing much larger numerators, it responds very sensitively to directional shifts. For instance, examinee 2 made equal numbers of positive and negative alterations between items 10-20 which produced a standardized change of zero. However, the unidirectional alterations taking place between items 30-40 of examinee 2 and throughout examinee 3 produce sometimes great shifts depending upon their quantity and the item difficulties associated with the measures.

Further, it is evident that such measure change statistics require the use of small subsets of items. When the same statistic was calculated across the entire test, very different and very difficult to interpret results were obtained.

Discussion

There are a variety of other possible approaches to detecting aberrant alteration patterns in the computer adaptive environment. Most such approaches suggest the use of Runs tests or a Delta statistic. One additional and intriguing possibility involves the investigation of "time on task". This approach explores the variation of time spent on each item (within items, before and after review) and across items (also before and after review).

This exploratory study does suggest that both regression and standardized measure change approaches may be viable techniques for the detection of response alteration patterns that are questionable. Much more study with the use of real life data is required as there are a number of questions still unanswered.

One of the primary concerns about each of these approaches is sensitivity. The objective is to detect questionable patterns which do not fit with the adaptive model being used. The level at which these diversions become significant is yet unknown. Clearly the standardized measure change approach is extremely sensitive. A few items altered in any one direction within a small group of items can produce major shifts. Summed across the test these may be very inflated. Yet, calculating across the entire test in a single measure, is not sufficient to detect the patterns at all. The sensitivity of the new fit statistic must be established so that misfit is neither inflated, nor understated. On this point, the regression approach appears to be less volatile.

Educational Importance

The ability to detect misfit is essential for all measurement. Thusfar, computer adaptive tests have relied on fit estimates better suited to traditional paper and pencil examinations. Since fit has been traditionally examined in a global manner, it will be important to understand fit on a subset level, and perhaps on an item by item level, in order to best assess fit in the new controlled environment. The implications of these discussions may extend to all types of examinations being given in the computer adaptive format. Although cheating has not been shown to plague the system, if its immediate detection is possible, then the concern over the perceived potential for cheating may be greatly and systematically reduced. Much more work is called for to refine the aspects of

both approaches discussed and to explore the other viable alternatives including those suggested.

Sources

Bergstrom, B. & Lunz, M.E. (1991). Effect of Review Among Ability Groups on Two Computer Adaptive Tests. Paper presented at the annual meeting of the NCME, Chicago, Ill.

Coombs, C.H. (1964). A Theory of Data. New York: John Wiley & Sons.

Cronbach, L.J. (1946). Response sets and test validity. Educational and Psychological Measurement, 6, 475-494.

Gershon, R. (1990). CAT ADMINISTRATOR [Computer program]. Chicago: Computer Adaptive Technologies.

Green, B.F., Bock, R.D., Humphreys, L.G., Linn, R.L., and Reckase, M.D. (1984). Technical guidelines for assessing computerized adaptive tests. Journal of Educational Measurement, 4, 241-261.

Guttman, L. (1944). A basis for scaling qualitative data. American Sociological Review, 9, 139-150.

Lord, F. M. (1983). Small N justifies Rasch model. In D.J. Weiss (Ed.), New Horizons in Testing (pp. 51-61). New York: Academic.

Lunz, M.E., Bergstrom, B. & Wright, B.D. (1991). The Effect of Review on Student Ability and Test Efficiency for Computer Adaptive Tests. Paper presented at the annual meeting of the NCME, Chicago, Ill.

Mead, R. (1976). Assessment of Fit of Data to the Rasch Model Through Analysis of Residuals. Ph.D. Dissertation, University of Chicago: Chicago, Ill.

Mead, R.J. (1979). Using the Rasch Model to Identify Person-Based Measurement Disturbances. Proceedings of the 1979 Computerized Adaptive Testing Conference, University of Minnesota.

Olsen, J.B., Maynes, D.D., Slawson, D., & Ho, K. (1986). Comparison and equating of paper-administered, computer administered and computerized adaptive tests of achievement. Paper presented at the meeting of the American Educational Research Association, San Francisco.

Rasch, G. (1960). Probabilistic Models for Some Intelligence and Attainment Tests. Chicago: University of Chicago Press.

Smith, R.M. (1982). Detecting Measurement Disturbances with the Rasch Model. Ph.D. Dissertation, University of Chicago: Chicago, Ill.

Stone, G.E. & Lunz, M.E. (1994). The effect of review on the psychometric characteristics of computerized adaptive tests. Applied Measurement in Education, 7,(3), pages 211-222.

Thurstone, L.L. (1928). The Measurement of Values. University of Chicago Press.

Wright, B.D. & Linacre, J.M. (1991). BIGSTEPS (Computer Program). Chicago: MESA Press.

Wright, B.D. & Stone, M. (1979). Best Test Design. Chicago, Ill.: MESA Press.

Table 1: Examinee Measures and Fit Statistics

	Examinee 1: No Alterations	Examinee 2: 8 Alterations (2 correct to incorrect 6 incorrect to correct)	Examinee 3: 9 Alterations (All incorrect to correct)
Measure Before Review	1.81	1.64	-0.44
SE Before Review	0.30	0.30	0.30
Infit MNSQ Before Review	0.97	0.99	1.00
Outfit MNSQ Before Review	0.99	0.99	1.03
Measure After Review	1.81	1.81	0.38
SE After Review	0.30	0.30	0.35
Infit MNSQ After Review	0.97	1.01	1.03
Outfit MNSQ After Review	0.99	1.00	1.06

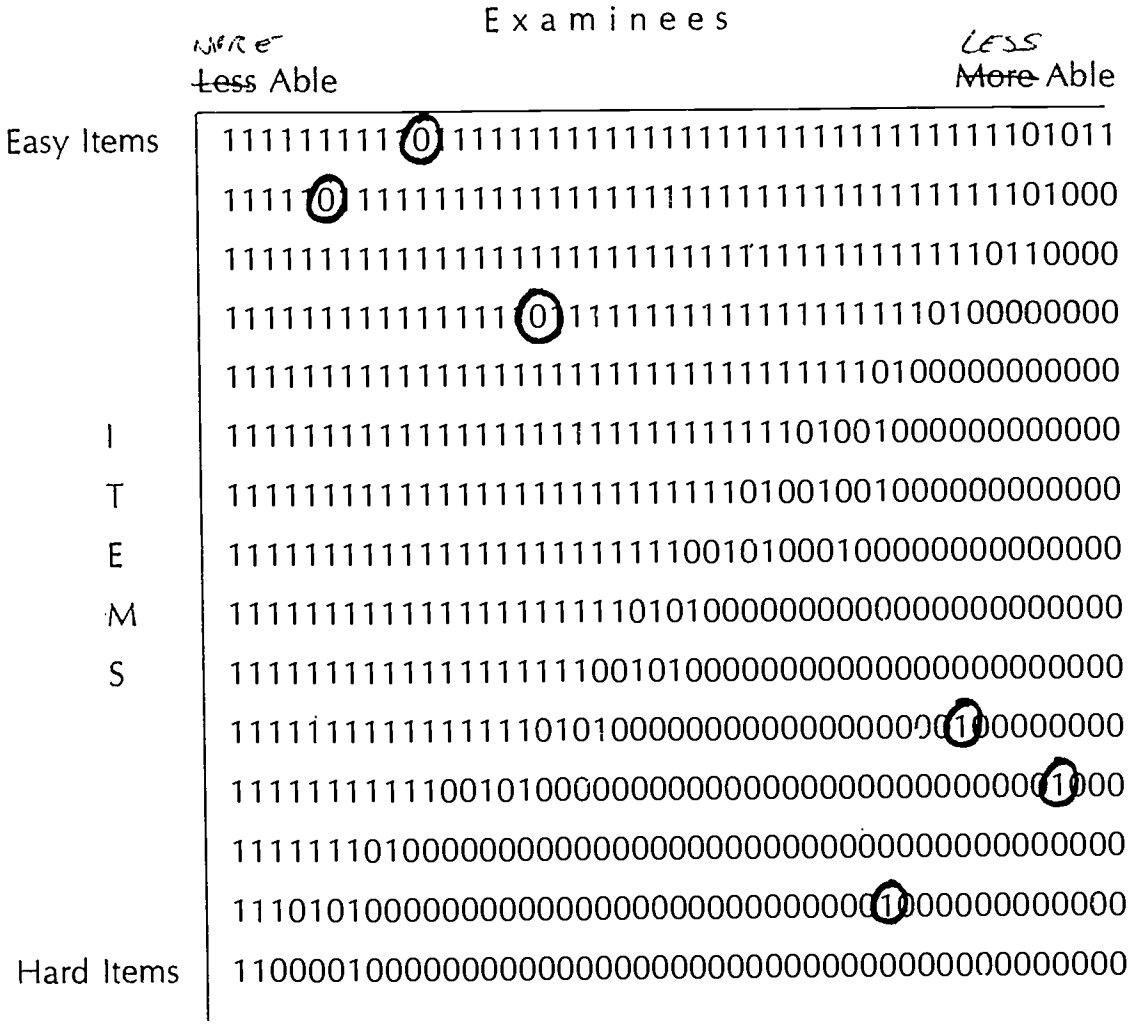
Table 2: Results of Regressions (after review measure with before review measure)

	Examinee 2	Examinee 3
R^2	0.930	0.032
F	621.84	0.050
F significance	< .001	0.829
T	24.94	-0.22
T significance	< .001	0.829

Table 3: Analysis of Measure Changes

	Examinee 2			Examinee 3		
Item	Before	After	Change	Before	After	Change
10	0.19	0.19	-	0.24	0.43*	+.19
20	0.29	0.29*	-	0.26	0.75*	+.49
30	0.02	0.02	-	0.36	0.54*	+.18
40	0.34	0.52*	+.18	0.00	0.21*	+.21
50	0.29	0.45*	+.16	0.00	0.00	-
Total	1.13	1.47	+.34	0.86	1.93	+1.07
Total Test (from 1-50)				Total Test (from 0-50)		
	0.42	0.29	-.13	0.00	0.05	+.05

Figure 1: Expected Guttman Response Pattern (Simulated data)



0 = Incorrect Response 1 = Correct Response

Figure 2:

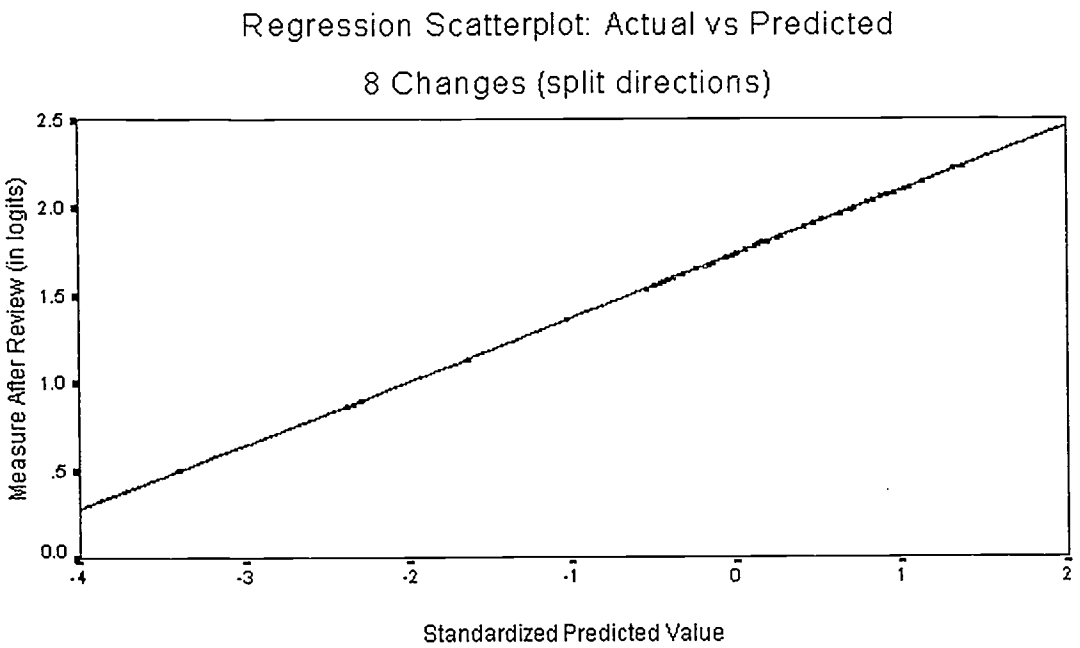

Regression Scatterplot: Actual vs Predicted
8 Changes (split directions)
Measure After Review (in logits)
2.5
2.0
1.5
1.0
.5
0.0
-4
-3
-2
-1
0
1
2
Standardized Predicted Value

Figure 3:

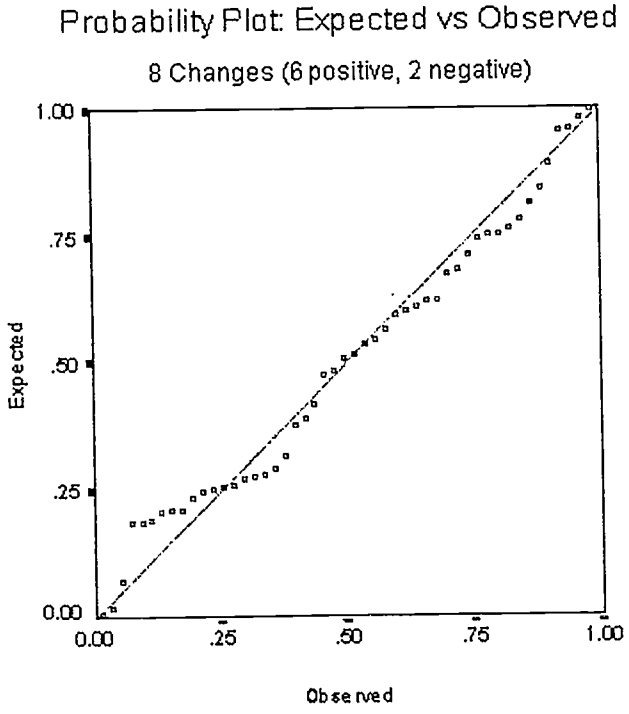

Probability Plot: Expected vs Observed
8 Changes (6 positive, 2 negative)
1.00
.75
.50
.25
0.00
Expected
0.00
.25
.50
.75
1.00
Observed

Figure 4:

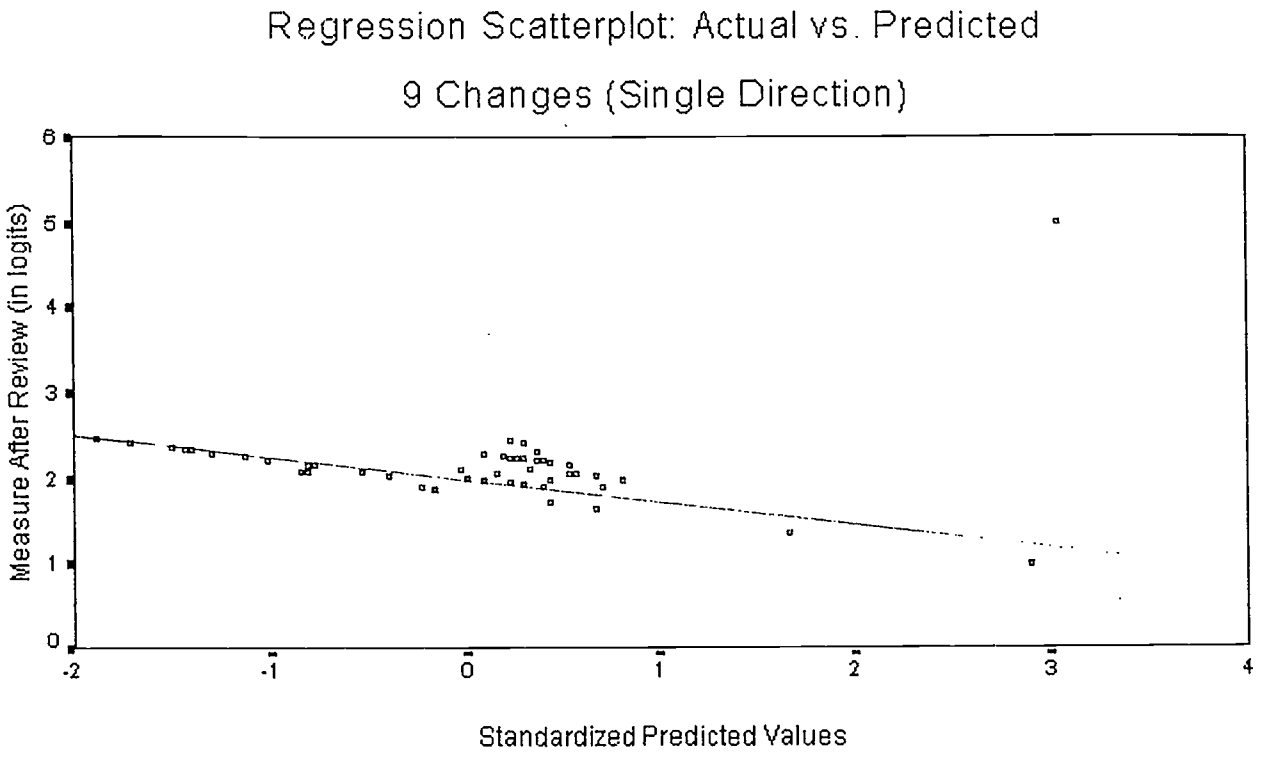
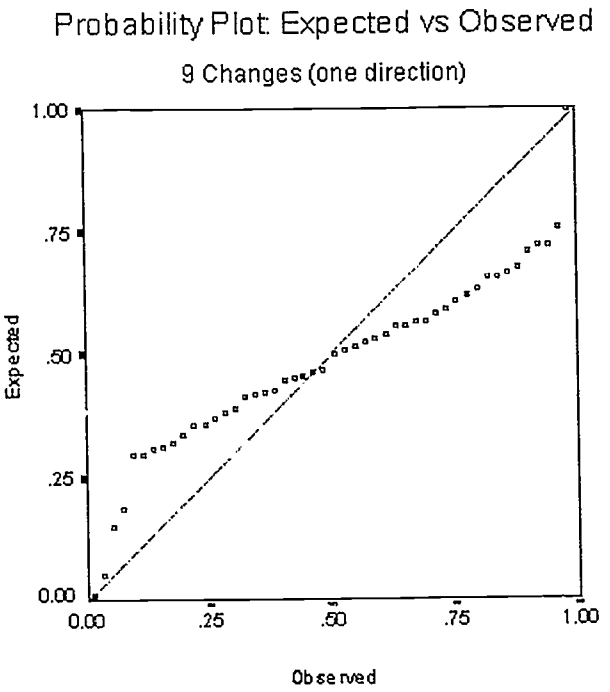

Regression Scatterplot: Actual vs. Predicted
9 Changes (Single Direction)
Measure After Review (in logits)
0
1
2
3
4
5
6
-2
-1
0
1
2
3
4
Standardized Predicted Values

Figure 5:


Probability Plot: Expected vs Observed
9 Changes (one direction)
Expected
Observed
1.00
.75
.50
.25
0.00
0.00
.25
.50
.75
1.00