Impact of Eliminating Anchor Items Flagged From Statistical Criteria on Test Score Classifications in Common Item Equating

Thakur Karkee
Seung Choi

CTB/McGraw-Hills

Paper Presented at the Annual Meeting of the American Educational Research Association, Montreal
April 15, 2005

Abstract

Proper maintenance of a scale established in the baseline year would assure the accurate estimation of growth in subsequent years. Scale maintenance is especially important when the state performance standards must be preserved for future administrations. To ensure proper maintenance of a scale, the selection of anchor items and evaluation of their performance in the succeeding administrations are crucial.

Under the common item equating design, anchor item selection is a critical process and is typically scrutinized against a set of guidelines for selection. Even if the guidelines for selecting anchor items are met for the anchor set, items may still perform differentially, for example, context effects, or changes in instructional emphasis, which might introduce additional dimensions, hence, differential examinee performance.

In this study, four statistical methods of evaluating anchor items are explored. The evaluation process used data from statewide assessments in two content areas. The results indicate that the Delta-Plot method which uses proportion correct (p-value) as input, differs from the methods (Lord's Chi-Square, Stocking and Lord's TCC inverse, and Raju's area minimization methods) that use Item Response Theory (IRT) item parameters as input in flagging differentially performing items. The results further show that the decisions of excluding or including anchor items impacts on students' test scores at both the student and group levels.
Keywords: item response theory (IRT), Equating, Anchor

Objective

The passage of ESEA (the Congressional Act of the President's "No Child Left Behind" initiative) requires that all elementary and secondary public schools set and keep a scale for the purpose of longitudinal data tracking. Many states are in the process of establishing new scales for the federally mandated tests and some are using previously established scales. In both cases, proper maintenance of the scale established in the baseline year would ensure the accurate estimation of growth in the subsequent years. Scale maintenance is especially important when the state performance standards must be preserved for future administrations.

Proper maintenance of a scale in a common-item non-equivalent groups design depends upon the anchor items. Guidelines for selecting anchor items have been established and are available. For example, the anchor set should be a miniature version of the total test in terms of adequate content representation, higher item-total score correlations, and spread of item difficulties. Similarly, the anchor items should not exhibit poor fit, or differential item functioning (DIF) for subgroups, and the exposure of anchor items must be controlled. Even if the guidelines are strictly met for selecting an anchor set, individual item may still perform differentially perhaps due to context effects, or instruction with different curricular emphases, which might introduce additional dimensions, hence, differential examinee performance (e.g. Miller and Linn, 1988; Bock, Muraki and Pfeiffenberger, 1988).

Statistical methods for flagging differentially performing anchor items use different parameters as input variables. They are briefly described in the next section. However, there is no unambiguous method for evaluating the magnitude of differential performance of anchor items during the equating phase. If an item is flagged for performing differentially across administrations from a statistical criterion, should the item be eliminated? What would be the impact if the elimination of an anchor item compromised the content representation? This study attempts to document anchor item evaluation methods and explore the impacts of the decisions made based on the statistical criteria on the equating transformations, score distributions, and the classification of students into different proficiency levels.

The significance of this study lies particularly in light of measuring average yearly progress on large-scale assessments. Under the current No Child Left Behind (NCLB) plan schools are held accountable for their student test scores. Practitioners must make decisions regarding the equating of alternate forms of tests. The nature of such decisions may vary according to the equating design. This study addresses one of the designs frequently used to deal with ability differences in the samples used to equate alternate forms of tests. When item characteristics vary from sample to sample, the construct being measured is called into question. A remedy might be to eliminate misbehaving anchor items from defining the relationship of a new form of the test to the existing scale. However, this decision may call into question whether the anchor items adequately represent a target test in terms of difficulty and content. In large-scale testing programs these target scales will benchmark the progress of cohorts of students as

agencies strive to demonstrate greater efficiencies in teaching. Since goals are being set for such increases in efficiency, some guidelines are needed whereby practitioners can decide whether a set of anchor items are serving the intended purposes of equating or not.

Theoretical Framework

Four statistical methods for evaluating anchor items include delta plot (Angoff, 1972; Dorans and Holland, 1993), iterative linking (Candell & Drasgow, 1988) using Stocking and Lord's (SL) (1983) test characteristic curve method, Lord's chi-square criterion (p. 223, Lord, 1980), and an iterative process using area minimization and the significance of areas (Raju & Arenson, 2002; Raju, 1990).

The delta-plot method relies only on the differences in the proportion correct value (p-value). For example, p-values of the anchor items based on the previous and current year's population will be calculated. The p-values will then be converted to z-scores that correspond to the $(1\text{-p})^{th}$ percentiles. For example, for a p-value of 0.90, the corresponding z-score will be the $(1\text{-}.90)^{th}$ percentile, which is -1.2816. A simple rule to identify outlier items that are functioning differentially between the two groups with respect to the level of difficulty is to draw perpendicular distance to the line of best fit. The fitted line is chosen so as to minimize the sum of squared perpendicular distances of the points to the line. The perpendicular distance is given by:

$$D = \frac{AZ_{old} - Z_{new} + B}{\sqrt{A^2 + 1}}$$

Where $A = \dfrac{(SD^2_{Znew} - SD^2_{Zold}) + \sqrt{(SD^2_{Znew} - SD^2_{Zold})^2 + 4r^2_{(Zold)(Znew)} SD^2_{Zold} SD^2_{Znew}}}{2r_{(Znew)(Zold)} SD_{Zold} SD_{Znew}}$

and B = Mean (Z_{new})-A*Mean (Z_{old}). The standard deviation (SD) of the perpendicular distance is given by:

$$SD_D = [(SD_{Znew} + SD_{Zold})/2] * \sqrt{1 - r_{(Zold)(Znew)}}$$

As a rule of thumb, any items lying more than two standard deviations of the distance away from the fitted line are flagged as outliers.

Stocking and Lord (1983) procedure, also called test characteristic curve (TCC) method, minimizes the mean squared difference between the two TCCs, one based on estimates from the previous calibration and the other on transformed estimates from the current calibration. Let $\hat{\psi}_j$ be the test characteristic curve based on estimates from previous calibration and $\hat{\psi}_j^*$ be the test characteristic curve based on transformed estimates from the current calibration.

$$\hat{\psi}_j = \hat{\psi}(\theta_j) = \sum_{i=1}^{n} P_i(\theta_j; a_i, b_i, c_i)$$

$$\hat{\psi}_j^* = \hat{\psi}(\theta_j) = \sum_{i=1}^{n} P_i(\theta_j; \frac{a_i}{M_1}, M_1 b_i + M_2, c_i)$$

The TCC method determines the scaling constants (M1 and M2) by minimizing the following quadratic loss function (F):

$$F = \frac{1}{N} \sum_{a=1}^{N} (\hat{\psi}_j - \hat{\psi}_j^*)^2$$

The differential item functioning was evaluated by examining previous (input) and transformed (estimated) item parameters. The item with absolute difference of parameters greater than two times the root mean square deviation was flagged. The

difference was also monitored by plotting input and estimated item parameters and flagging the one with substantially larger differences.

Lord's χ^2 criterion involves significance testing of both item difficulty and discrimination parameters simultaneously for each item and evaluating the results based on the chi-square distribution table (see Divgi, 1985 & Lord, 1980 for detail). If the null hypothesis that the item difficulty and discrimination parameters are equal the χ^2 follows chi-square distribution with 2 degrees of freedom.

The area minimization method (Raju & Arenson, 2002, & Raju, 1990) determines scaling constants by minimizing the area between the two ICCs. The exact unsigned area (EUA) between two ICCs based on the 2-PL model is given by:

$$f(AB) = H_1^2 + H_2^2 + \ldots\ldots\ldots\ldots + H_n^2, \text{ Item i=1}\ldots.\text{n}$$

Where

$$H_i = \frac{2(Aa_{1i} - a_{2i})}{Da_{1i}a_{2i}} \ln\left\{1 + \exp\left[\frac{Da_{1i}a_{2i}}{Aa_{1i} - a_{2i}}(b_{1i} - Ab_{2i} - B)\right]\right\} - (b_{1i} - Ab_{2i} - B)$$

Substituting H_i on $f(AB)$ and deriving partial derivatives with respect to A and B using Broyden-Fletcher-Goldfarb-Shanno (BFGC) algorithm (Press, Teukolsky, Vetterling, & Flannery (2002) yields solutions for A and B. The evaluation criteria includes plotting of input and estimated *a* and *b* parameters and flagging the item that lies far from the fitted line. The anchor item is flagged if the absolute difference between the input and estimated *b* parameter is greater than 0.5 where the item parameters are in 0/1 metric.

While dropping an anchor item flagged from the statistical criteria has its simplicity, this option may change the content coverage and equating constants, shift scale score distributions, and affect the classification of students by moving them into different proficiency levels. This is especially true when the number of anchor items in the set is small. It may be needed to include flagged items in the test for an adequate coverage of the content standards, even though the quality of the anchor items is dubious.

Data Source

Algebra and Reading/Language arts assessments from a large-scale standards-based statewide operational assessment provided the data for this study. The test configuration is shown in Table 1. Algebra is measured by three content standards: number sense and algebraic operations, relations and functions, and data analysis with a total of 53 items, all multiple-choice (see Table 1A). Reading/Language Arts is measured by six content standards: vocabulary, comprehension, literature, research and information, grammar usage and mechanics, and writing process with a total of 62 items of which 60 items are multiple-choice and two are constructed-response (CR) items. Each CR item in Reading/Language Arts carried six score points totaling to 72 score points (see Table 1B) for the test.

The anchor item evaluation was conducted in two phases. In the first phase, the four anchor item evaluation methods were compared based on the number of items flagged for performing differentially. The item characteristics of the flagged items were further studied. In the second phase, the impact of keeping or dropping the problematic

items from the anchor set was evaluated based on the change in content coverage, correlation coefficients between the input and estimated anchor item parameters, scale score distribution, and classification of students into different proficiency levels. The data consisted of 38,174 valid responses on the 53 items for Algebra and 33,729 valid responses on the 62 items for Reading/Language Arts.

Methods

The three-parameter logistic model (Lord, 1980) was used for scaling MC items and the two-parameter partial credit model (Yen, 1981) was used for the CR items. The PARDUX microcomputer program (Burket, 2002) was used for calibration and scaling. It simultaneously calibrates the MC and CR items using the Marginal Maximum Likelihood Estimation technique for both item parameters and person ability estimation. The program constrains the mean and standard deviation (SD) of the examinee ability distribution to 0 and 1, respectively, during the item parameter estimation process to obtain model identification.

For equating and parameter comparison, a program was written in S-PLUS for the Lord's Chi-square, Raju's Area Minimization, and Delta-Plot methods. Stocking and Lord's TCC method was implemented in PARDUX.

Results/Conclusions

A summary of flagging criteria for different anchor item evaluation methods and flagged items are listed in Table 2. It shows that the methods that utilized IRT item

parameters as input (Raju's Area Minimization, Lord's Chi-Square, and Stocking and Lord's Inverse TCC methods) consistently flagged the same items (items 11 and 49 in Algebra, and items 16 and 19 in Reading/Language Arts) for performing differentially across two administrations, whereas the method that utilized proportion correct value (Delta-Plot method) did not flag any item. The details of *a* and *b* plots under Raju's area minimization and Stocking and Lord's TCC methods, Chi-Square and associated p-values under Lord's Chi-Square, and perpendicular distance from the fitted line under Delta-plot methods are shown in Appendix A.

Content Representation

The content coverage by the anchor set can be compromised as a result of dropping the two anchor items flagged by the three methods (see Tables 1A and 1B). For example, there was as much as 5% change in content coverage by dropping items 11 and 49 from the Algebra anchor set. Similarly, the "vocabulary" content standard was not represented at all and the "grammar/usage and mechanics" was over-represented by 11% from dropping items 16 and 19 from the anchor set in the Reading/Language Arts test.

Input and Equated Item Parameters Correlation

Despite the under- and over-representation of the content coverage, dropping the flagged items from the anchor set increased the correlation between the input and equated item parameters in both content areas (Table 3). For Algebra, the Pearson correlation coefficient for discrimination parameter increased from .84 to .85, location parameter from .88 to .96, psudo-guessing parameter from .74 to .93, and p-value from .97 to .98.

Similarly, the correlation coefficients for Reading/Language Arts increased from .89 to .95, location parameter from .92 to .96, and guessing parameter from .47 to .71. The correlation between the p-values for the Reading/Language Arts before and after dropping the flagged items, however, did not change.

Item/Test Characteristics

The item characteristic curves for the flagged items and test characteristics curves for input and estimated anchor sets, and whole test between the two administrations for Algebra are shown in Figures 1a-1e and for Reading/Language Arts in Figures 2a-2e. These figures show that the item characteristics for the anchor items in question changed considerably between two administrations. The item 11 of Algebra was relatively high discriminating for the reference group (previous cohort) whereas item 49 was relatively high discriminating for the focal group (current cohort). Similarly, both items 16 and 19 of Reading/Language Arts were relatively high discriminating for the focal group (current cohort). Note that the p-value increased for the current administration for Algebra items and decreased for Reading/Language Arts items. Despite the two items flagged for performing differentially in each of the two content areas, the overall test characteristic curves for the input and estimated anchor sets were reasonably close to each other indicating that the anchor sets overall performed similarly between the two administrations.

Scale Score Distribution and Proficiency Classification

The changes to the mean scale score and classification of students into different proficiency levels are considered as the indices of the impact of dropping the two anchor items at student and group levels. The results are shown in Table 4. It indicates that the scale score mean decreased by approximately 2.4 scale score points by dropping the two anchor items from the anchor set in Algebra. The trend, however, reversed for Reading/Language Arts with the scale score mean increased by approximately 1.4 scale score points. The proficiency level classifications did not change significantly. The increasing or decreasing trends, however, are reflected in the proficiency level classifications as well. For example, dropping the two flagged items decreased the percentage of students at or above "Satisfactory" by approximately 0.5% in Algebra and increased by approximately 0.9% in Reading/Language Arts. In Algebra, about 1.5% more students classified as "Unsatisfactory" whereas in Reading/Language Arts about 1.4% less students classified as "Unsatisfactory" by dropping the two flagged anchor items from the anchor sets.

Summary and Discussion

This study examined four anchor item evaluation methods using two different content areas. Note that the statistical evaluation of anchor items is a post-hoc process to check items that may have performed differentially between two administrations after placing them on a common scale. Anchor items are selected through a rigorous process, in which they are evaluated on the basis of individual item characteristics as well as the anchor set as a whole. For example, an anchor set should reflect the test blueprint in terms of content representation, range of p-values as well as item location, and point

biserials. Similarly, the anchor items should be free from differential item functioning and poor model fit. In order to minimize the context effect, anchor items are placed within the same relative positions where they appeared in the previous administration, and the format of the items is strictly maintained. Despite these efforts to maintain the anchor item integrity, some anchor items still performed differentially overtime, perhaps due to excessive exposure or shift of curricular emphasis.

Results from this study showed that different items may be flagged by different statistical methods. Specifically, the method based on p-values (i.e., the Delta-Plot method) produced different results, compared to the methods that use item parameters as input (for example, Stocking and Lord and Lord's Chi-Square method and Raju's area minimization method). Note that the p-values are sample dependent and the item parameters are, theoretically, sample invariant. From that perspective, the results indicate that the flagged items did not perform differentially between the two cohorts but the item characteristics changed significantly. Of course, this assumes that the conventional rules of thumb employed by the four methods are set with comparable rigor or sensitivity. A further investigation regarding the comparability of the four procedures with respect to their power is warranted.

One of the possible outcomes, although not observed in this study, is that the same anchor items are flagged by all four methods. That means, the flagged item performed differentially between the two cohorts and also the item characteristic changed significantly. This situation may result in or justify removing the item from the anchor

set. However, if removing the flagged anchor items overly violates the content representation one should abstain from dropping it.

The results from this study further illustrated that dropping flagged items from the anchor set has an impact on the equating transformation and the scale scores that students receive. For example, the mean scale score would increase if we dropped the anchor items with lower p-values compared to the reference administration. Although Psychometricians believe that dropping or keeping of a flagged item should be based solely on the performance of the anchor items but not on the impact on student scores, there is no lack of other perspectives that are more subtle and judgmental. In order to minimize the risk of inadvertently changing the content representation of an anchor set or causing unexpected results, it is critical to establish a clear and stringent guideline for selecting anchor items and a priori criteria for flagging and dropping anchor items.

As described in the introduction section, one of the purposes of this study was to document anchor evaluation methods. The results indicated that the three anchor item evaluation methods that utilized IRT item parameters (Raju's Area Minimization, Lord's Chisquare, and Stocking and Lord's TCC methods) performed similarly in identifying the differentially performing items between the two administrations. These methods consistently flagged the same items for performing differentially for both content areas. The method that utilized classical item statistics (Delta-Method) was not sensitive enough to detect the differences.

One of the limitations of this study is the small number of tests examined—only two test forms were investigated. In order to generalize the findings from this study, further research is warranted with a larger number of test administrations. A different number of anchor items may also need to be examined, because the impact of suppressing anchor items will be inversely related to the number of items in the original anchor set.

References

Angoff, W. H. (1972, September). A technique for the investigation of cultural differences. Paper presented at the annual meeting of the American Psychological Association, Honolulu.

Bock, R. D., Muraki, E., & Pfeiffenberger, W. (1988). Item pool maintenance in the presence of item parameter drift. *Journal of Educational Measurement, 25*(4), 275-285.

Burket, G. (2002). PARDUX [Computer program]. Unpublished. Monterey, CA: CTB McGraw-Hill.

Candell, G. L., & Drasgow, F. (1988). An iterative procedure for linking metrics and assessing item bias in item response theory. *Applied Psychological Measurement, 12*, 153-260.

Divigi, D. R. (1985). A minimum chi-square method for developing a common metric in item response theory. *Applied Psychological Measurement, 9*, 413-415.

Dorans, N. J., & Holland, P. W. (1993). DIF Detection and Description: Mantel-Haenszel and Standardization. In P. W. Holland, and H. Wainer (Eds.), *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrance Erlbaum Associates.

Miller, A. D., & Linn, R. L. (1988). Invariance of item characteristic functions with variations in instructional coverage. *Journal of Educational Measurement, 25*(3), 205-219.

Press, W. H., Teukolsky, S. A., & Vetterling, W. T., & Flannery, B. P., (2002). *Numerical recipes in C++ : The art of scientific computing*. Cambridge, England : Cambridge University Press.

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika, 53*, 495-502.

Raju, N. S., & Arenson, E. (2002). Developing a common metric in item response theory: An area-minimization approach. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*, 201-210.

Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement, 5*, 245-262.

Table 1A. Number of Items and Percentage for Total Test and Anchor Set, Algebra

Test	CS1	CS2	CS3	Total
Number of Items	10	35	8	53
%	19%	66%	15%	100%
Anchor				
Number of Items	4	13	2	19
%	21%	68%	11%	100%
Modified by dropping items 11 and 49				
Number of Items	4	11	2	17
%	24%	65%	12%	100%

Content Standard 1 (CS1) - Number Sense and Algebraic Operations
Content Standard 2 (CS2) - Relations and Functions
Content Standard 3 (CS3) - Data Analysis

Table 1B. Number of Items and Percentage for Total Test and Anchor Set, Reading/Language Arts

Test	RL1.0	RL2.0	RL3.0	RL4.0	WG3.0	WG1.0/2.0	Total
Number of Points	6	19	15	4	16	12	72
%	8%	26%	21%	6%	22%	17%	
Anchor							
Number of Items	1	7	4	2	6	0	20
%	5%	35%	20%	10%	30%	0%	
Modified by dropping items 16 and 19							
Number of Items	0	6	4	2	6	0	18
%	0%	33%	22%	11%	33%	0%	

Reading/Language Arts (RL 1.0) = Vocabulary
Reading/Language Arts (RL 2.0) = Comprehension
Reading/Language Arts (RL 3.0) = Literature
Reading/Language Arts (RL 4.0) = Research and Information
Writing/Grammar (WG 3.0) = Grammar/Usage and Mechanics
Writing/Grammar (W/G 1.0/2.0) = Writing Process

Table 2. Flagging Criterion for Different Methods and Flagged Items

Methods	Flagging Criterion	Items Flagged	
		Algebra	Reading/Language Arts
Area Minimization	Absolute difference of location parameter >.5	11, 49	16, 19
Lord's Chi-Square	χ^2 > chi-square at 2 degrees of freedom at p<.05	11, 49	16, 19
Delta-Plot	Perpendicular distance > 2SD from the fitted line	None	None
Stocking and Lord	Absolute difference of b parameters greater than 2 times root mean square deviation	11, 49	16, 19

Table 3. Corrélation coefficients, r

Parameters	Algebra		Reading/Language Arts	
	r (19 Anchor items)	r (17 Anchor Items)	r (20 Anchor items)	r (18 Anchor Items)
Discrimination (a)	0.84	0.85	0.89	0.95
Location (b)	0.88	0.96	0.92	0.96
Pseudo-Guessing (c)	0.74	0.93	0.47	0.71
Difficulty (p)	0.97	0.98	0.98	0.98

Table 4. Scale Score Descriptive Statistics and Performance Level Distribution

Anchor Set	Unsatisfactory	Limited Knowledge	Satisfactory	Advanced	N	Mean	SD
Algebra							
19	29.5%	49.1%	12.2%	9.2%	38174	644.6	70.6
17	31.0%	48.1%	11.8%	9.1%	38174	642.2	71.7
Reading/Language Arts							
20	23.7%	20.0%	28.6%	27.8%	33729	703.6	70.2
18	22.3%	20.4%	29.7%	27.6%	33729	705.0	67.5

Figure 1: Item Charactersitic Curves for the Flagged Items and Test Characteristic Curves for Input and Estimated Anchor Sets and Whole Test, Algebra I

Figure 1a: Item 11 Focal Group

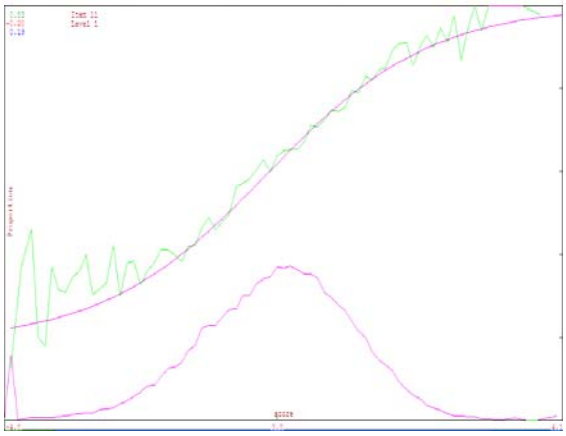


Figure 1b: Item 11 Reference Group

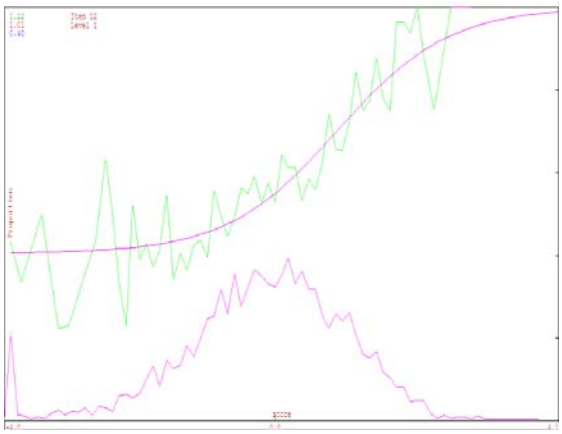


Figure 1c: Item 49 Focal Group

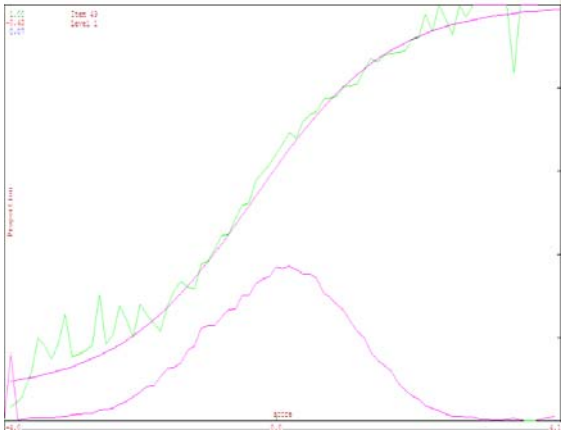


Figure 1d: Item 49 Reference Group

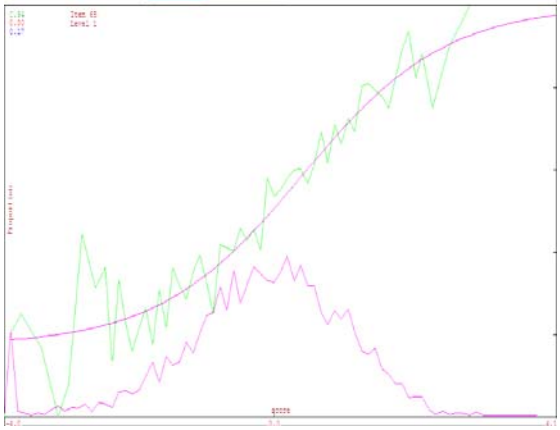


Figure 1e: Input and Estimated Anchor Sets, and Whole Test TCCs

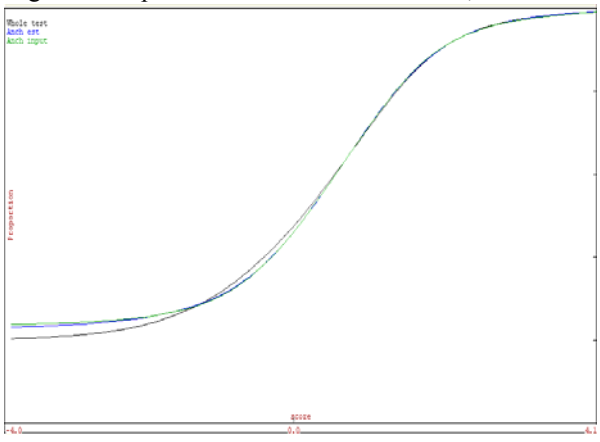


Figure 2: Item Characteristic Curves for the Flagged Items and Test Characteristic Curves for Input and Estimated Anchor Sets and Whole Test, English II

Figure 2a: Item 16 Focal Group

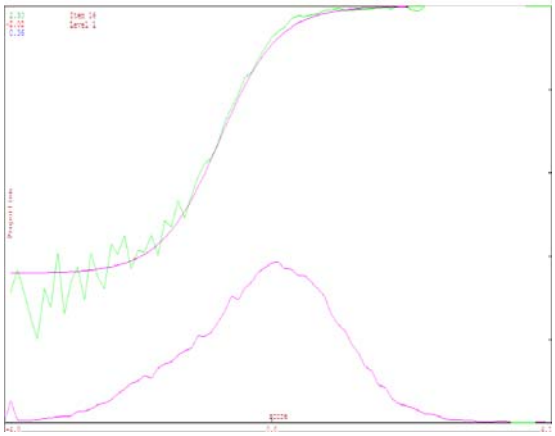


Figure 2b: Item 16 Reference Group

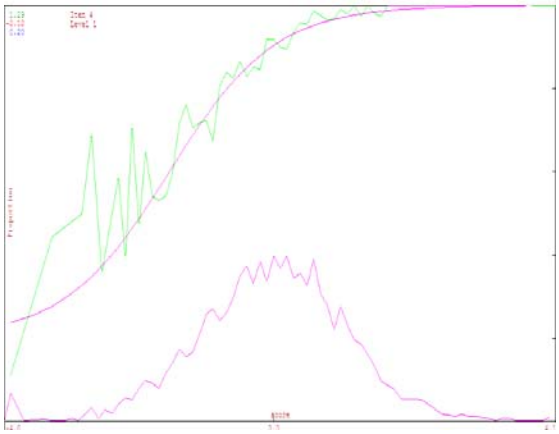


Figure 2c: Item 19 Focal Group

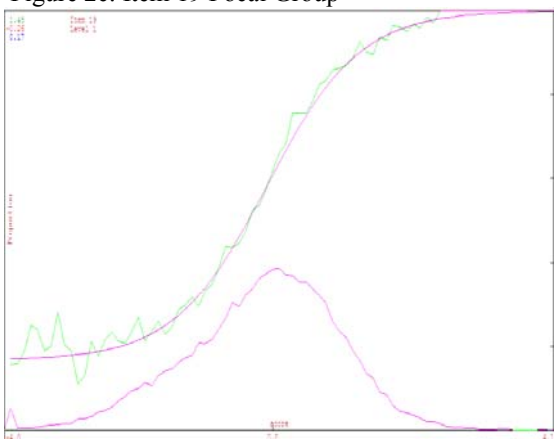


Figure 2d: Item 19 Reference Group

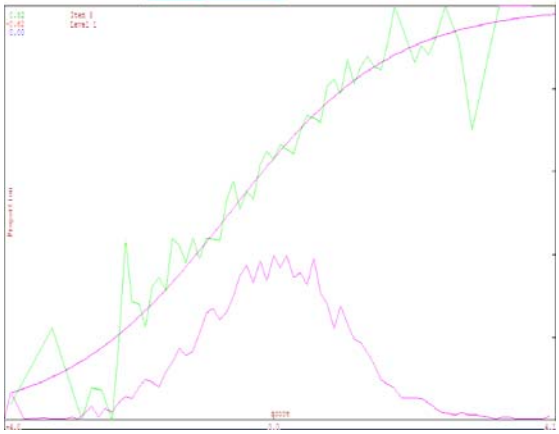
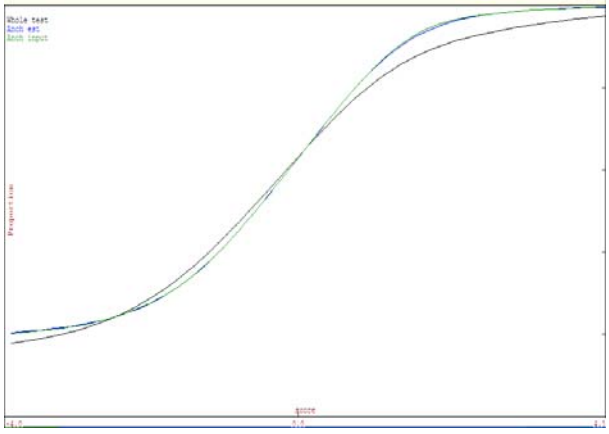


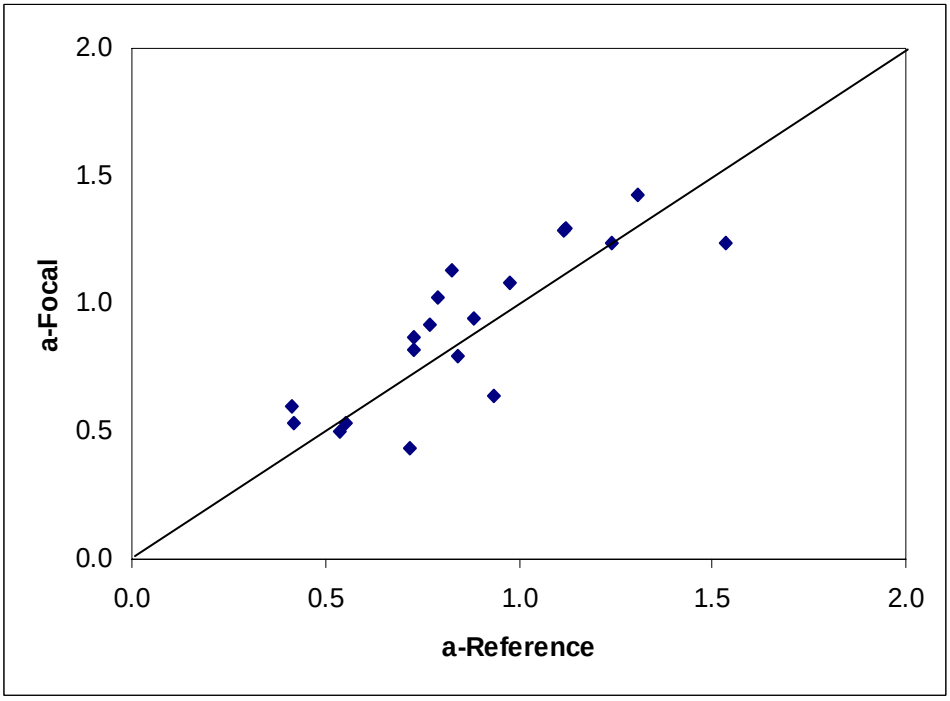
Figure 2e: Input and Estimated Anchor Sets, and Whole Test TCCs



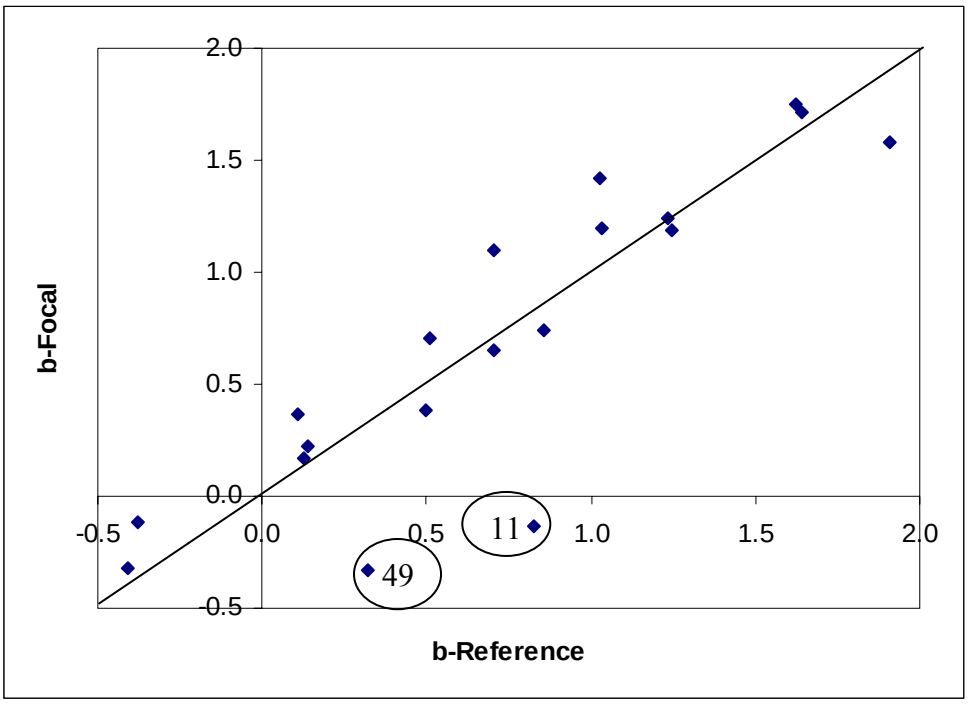
Appendix A. Flagged Items and Their Characteristics

Area Minimization Method

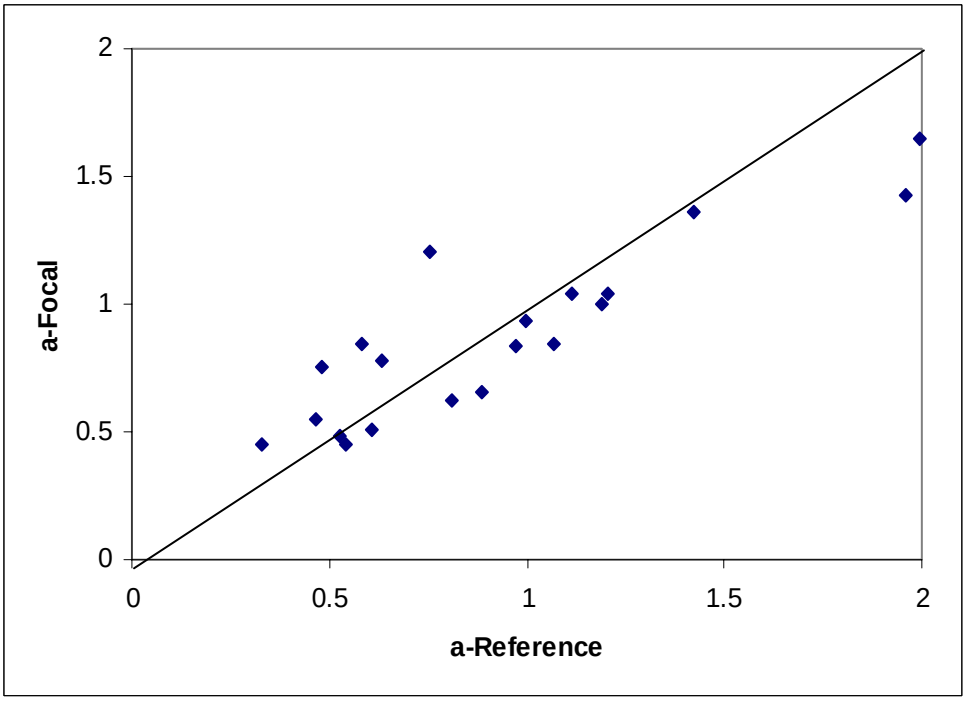
A-Plot: Algebra



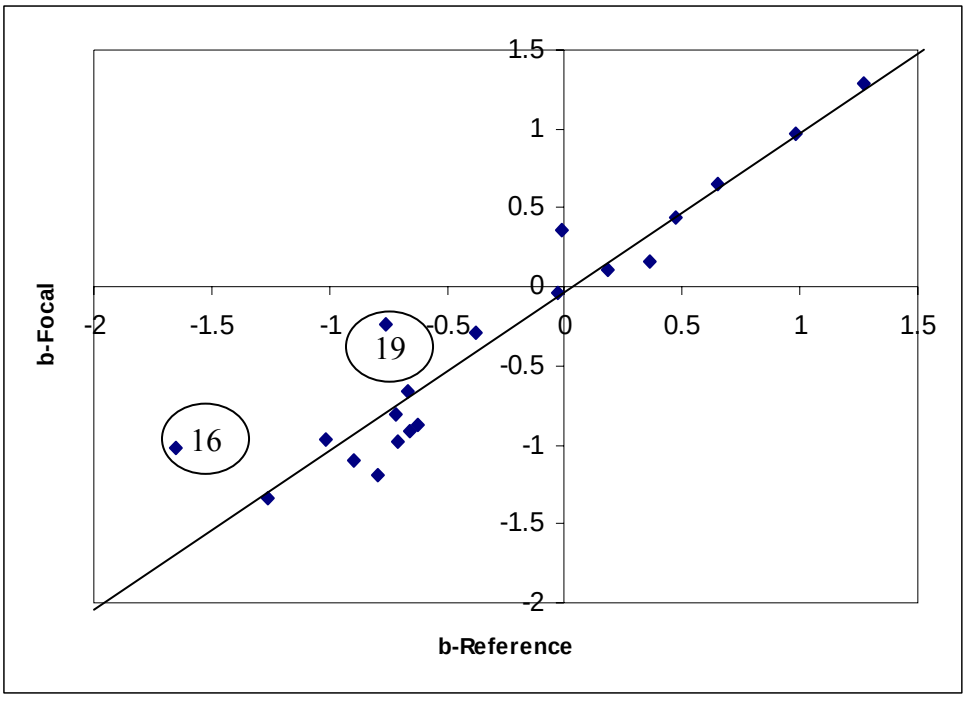
B-Plot: Algebra



A-Plot: Reading/Language Arts



B-Plot: Reading/Language Arts



Lord's ChiSquare

Table 5. Lord's Chi-square and Associated P-Values for Algebra and Reading/Language Arts

Algebra			Reading/Language Arts		
Item No.	**Chi-Square**	**P-Value**	**Item No.**	**Chi-Square**	**P-Value**
6	0.66	0.72	5	0.41	0.81
8	0.78	0.68	6	0.40	0.82
9	3.02	0.22	7	0.36	0.84
11	12.39	0.00	8	0.60	0.74
12	4.42	0.11	9	0.02	0.99
13	0.25	0.88	10	1.31	0.52
17	0.57	0.75	11	1.62	0.44
20	3.77	0.15	12	0.09	0.96
21	1.28	0.53	15	1.86	0.40
23	0.63	0.73	16	17.70	0.00
24	0.33	0.85	17	6.21	0.04
25	0.01	1.00	18	1.38	0.50
27	0.99	0.61	19	18.45	0.00
28	0.74	0.69	20	4.06	0.13
29	0.02	0.99	45	0.27	0.87
31	2.04	0.36	46	3.51	0.17
37	4.70	0.10	47	2.07	0.36
38	0.39	0.82	48	4.33	0.11
49	8.31	0.02	49	1.48	0.48
			50	3.45	0.18

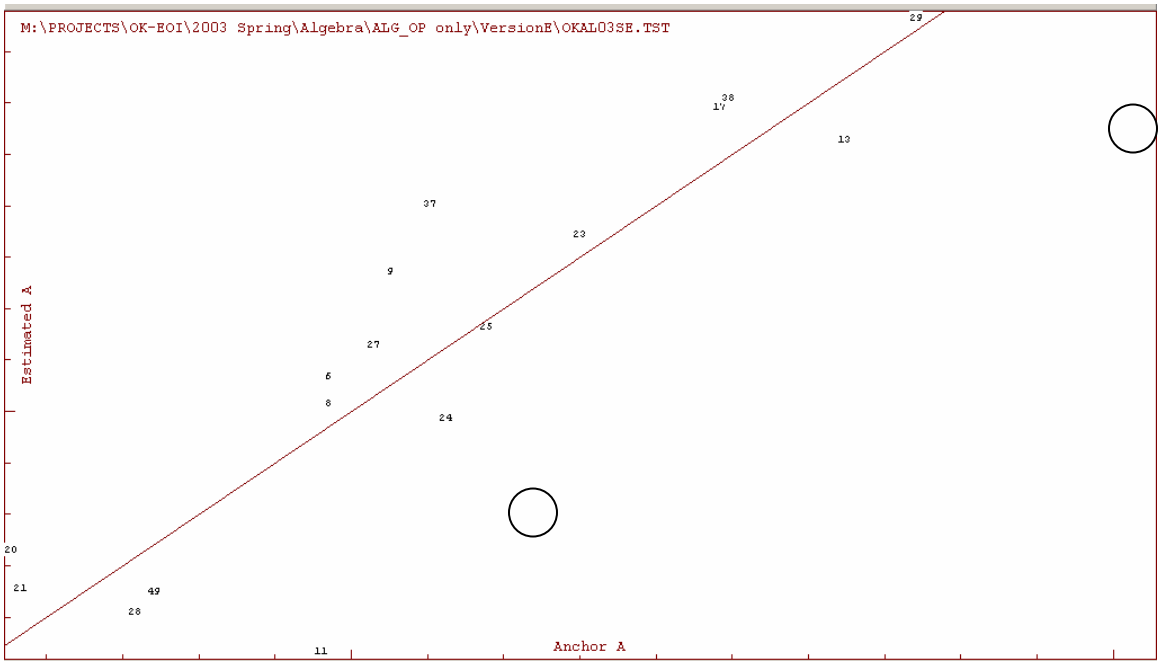
Delta-Plot Mehtod

Table 6. Perpendicular Distance From the Line of Best Fit Under Delta-Plot Method for Algebra and Reading/Language Arts

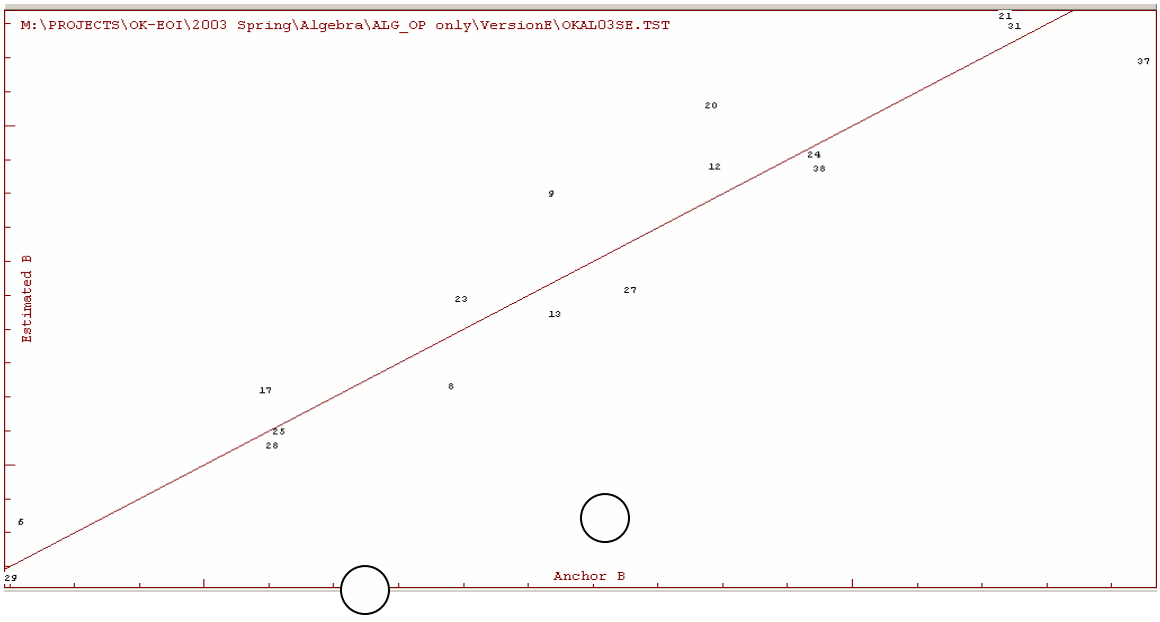
Algebra			Reading/Language Arts		
Item	**Perpendicular Distance from the line of best fit**	**Flag (2*SD)=.11**	**Item**	**Perpendicular Distance from the line of best fit**	**Flag (2*SD)= 0.14**
6	-0.02		5	0.00	
8	0.05		6	0.03	
9	-0.09		7	0.03	
11	0.03		8	0.06	
12	-0.07		9	0.02	
13	0.07		10	-0.01	
17	-0.03		11	0.06	
20	-0.10		12	0.06	
21	-0.02		15	-0.13	
23	-0.02		16	-0.11	
24	-0.01		17	-0.13	
25	0.01		18	-0.06	
27	-0.02		19	-0.08	
28	-0.02		20	-0.01	
29	0.03		45	0.05	
31	0.01		46	0.01	
37	0.04		47	0.08	
38	0.06		48	0.10	
49	0.11		49	0.01	
SD	**.054**		50	0.03	
			SD	**.069**	

Stocking and Lord's TCC Method

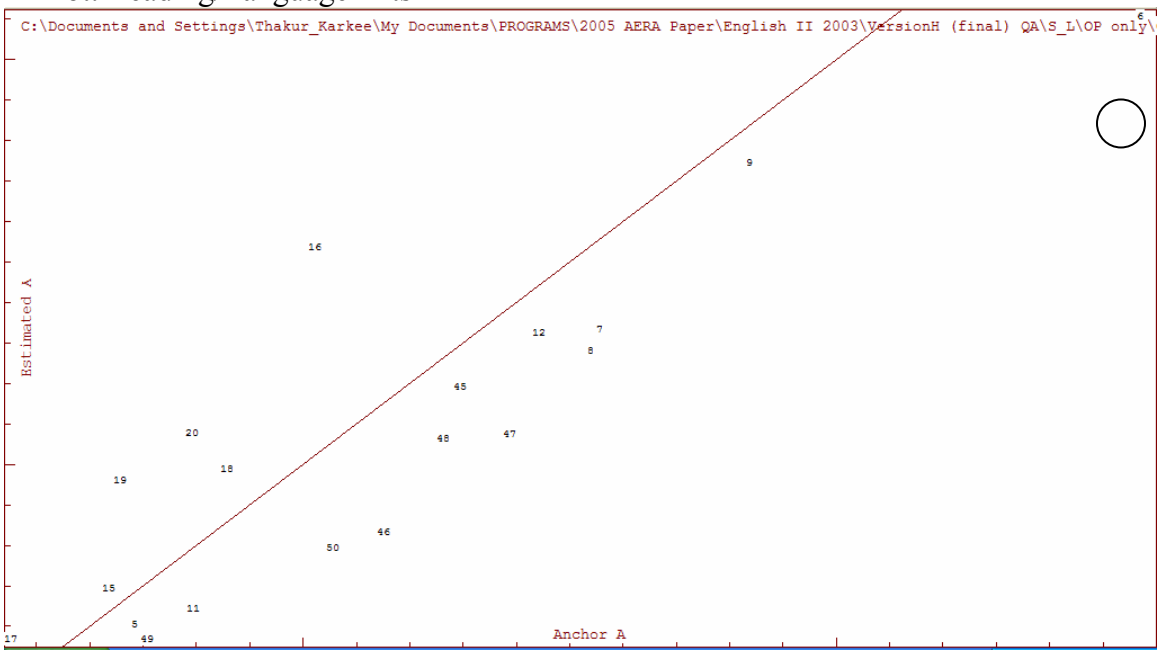
A-Plot: Algebra



B-Plot: Algebra



A-Plot: Reading/Language Arts



B-Plot: Reading/Language Arts