DOCUMENT RESUME

ED 359 266 TM 020 111

AUTHOR Du Bose, Pansy; Kromrey, Jeffrey D.

TITLE An Empirical Investigation of Equating Stability in a Single and a Double Linkage Design with Small Sample Sizes Using Angoff Model IV.

PUB DATE Apr 93

NOTE 48p.; Paper presented at the Annual Meeting of the American Educational Research Association (Atlanta, GA, April 12-16, 1993).

PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.

DESCRIPTORS Art Education; Comparative Testing; Computer Simulation; *Equated Scores; Graphs; Hearing Impairments; *Licensing Examinations (Professions); Mathematical Models; Monte Carlo Methods; Raw Scores; *Sample Size; Scoring; *Statistical Bias; Teacher Certification; *Test Items

IDENTIFIERS *Angoff Methods; Bootstrap Methods; Empirical Research; Linear Equating Method; *Linkage; Test Equivalence

ABSTRACT

Empirical evidence is presented of the relative efficiency of two potential linkage plans to be used when equivalent test forms are being administered. Equating is a process by which scores on one form of a test are converted to scores on another form of the same test. A Monte Carlo study was conducted to examine equating stability and statistical bias in a single and double linkage plan in small samples. Small random samples of 25, 50, and 100 were drawn with replacement from archival test data files representing Form B (base form), Form N (next form) and Form C (current form) pseudo-populations. Test data from two teacher certification subject area tests, Art Education and Hearing Impaired, both K-12 were used. Using the Angoff Model IV non-equivalent linear equating model, an indirect link, a direct link, and the average of the two links, equating equations were computed for each pair of samples at each sample size per subject area examination. Stability of the equating plans was evaluated by calculating the bootstrap standard errors of equating. Results indicate that the direct linkage design is more stable across raw score points, equating bias for direct linkage is trivial, and equating bias is quite large for the indirect linkage design. The direct linkage design is recommended for use with small sample sizes. Two tables and 13 figures illustrate the analyses. (Contains 12 references.) (SLD)

An Empirical Investigation of Equating Stability in a Single and a Double Linkage Design with Small Sample Sizes Using Angoff Model IV

Pansy Du Bose
Institute for Instructional Research and Practice
University of South Florida

Jeffrey D. Kromrey
Department of Educational Measurement and Research
University of South Florida

Paper presented at the Annual Meeting of the American Educational Research Association, April 12-16, 1993, Atlanta, GA

An Empirical Investigation of Equating Stability in a Single and a Double Linkage Design with Small Sample Sizes Using Angoff Model IV

Introduction

Test developers have several options from which to choose when they establish test equating strategies or linkage plans. The appropriate choice of strategies will help minimize equating errors and equating bias and stabilize the equating functions. Although test equating linkage plans are available, little empirical evidence is currently available to guide researchers and developers in their selection of such options.

According to Brennan and Kolen (1987), equating error that accumulates over multiple equatings in an equating linkage plan has not been extensively explored in the literature. What is known is that the degree of confidence in the stability of equating is inversely related to the number of equatings needed to progress from the new form to the initial base form (Kolen & Brennan, 1987).

In several equating models, the underlying assumptions of the models do not directly address the form or forms to which an anchor form was itself equated. The way in which equating models address this phenomenon is through the transitive property of equating (Kolen and Brennan, 1987). That is, if Form B is equated to Form N, and Form N is equated to Form C, then Form B is equated to Form C.

The stability of the links in an equating linkage plan can be analyzed through: (a) a single link design and (b) a double link design. Under Angoff Model IV linear equating design, two single links can be compared (Form C linked to Form B and Form N linked to Form B). If the two links yield similar results, then there is evidence of equating agreement (Kolen & Brennan, 1987; Cope, 1987). In double linking, one indirect and one direct C-B link is established: (a) Form C is indirectly linked to Form B through Form N, and (b) Form C is linked directly to Form B (See Figure 1). In most double linkage situations, the final equating equation for the C-B link is established by averaging the indirect and direct link parameter estimates.

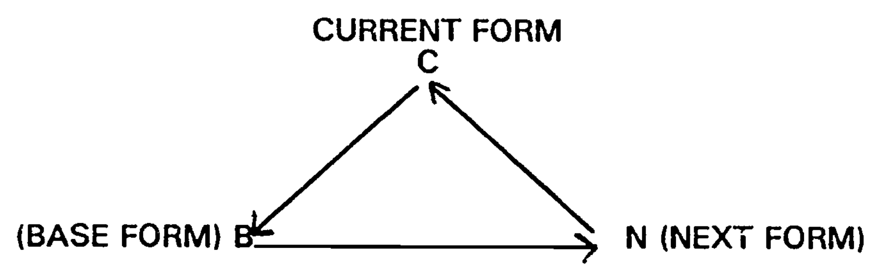

Figure 1. Linking Design Structure

The systematic study of test equating is a recent phenomenon in educational measurement and the knowledge of many aspects of equating is incomplete. In an effort to add to the knowledge base on test equating, an empirical research study was conducted to investigate a single and a double linkage equating designs using Angoff Model IV with small sample sizes.

A test-equating linkage plan is especially important i educational testing programs where equivalent test forms are being administered. Test developers and administrators must select from different equating data collection designs as well as from different equating linkage plans. This research study presents to test developers and researchers empirical evidence of the relative efficiency of two potential linkage plans.

Literature Review

Equating

In the construction of parallel test forms, the forms are assembled to be approximately equivalent in content, format and difficulty (Angoff, 1971). Since parallel test forms are not exact replicas and are only approximately equivalent, they can be made statistically equivalent by using equating procedures. Equating procedures are used to adjust for differences in test form difficulty, not test form content.

Equating is a process by which scores on one form of a test are converted to scores on another form of the same test. This process allows comparable comparisons to be made across examinee groups regardless of the test form administered. An unfair comparison would occur if the raw score of an examinee who by chance took a more difficult form of a test was compared to that of an examinee who by chance took an easier form of the same test.

The data collection design that is the focus of this research study is the common-item-non-equivalent groups design of linear equating. The common-item-non-equivalent groups design is a linear equating method by which a common set of test items, called an anchor test, is incorporated into the total test, either internally or externally (Budescu, 1985; Angoff, 1971). In this design, one form of a text (Form B) is administered to one group of examinees A, a second form (Form N) is given to a second group of examinees, B, and a common-item test (U) is administered to both groups (T). The common-item test (also called the anchor test) should be administered in the same order to both groups. The order of the anchor test should be maintained so that scores on the anchor test or on the old and new forms are affected in the same way by learning, fatigue, and practice effects (Petersen et al., 1989).

Scores on the anchor test are used to estimate the performance of the total group of examinees on both forms of the test, thus simulating by statistical methods, the situation in which the same group of examinees takes both forms of the test (Petersen, Kolen, & Hoover, 1989). Ideally, the anchor test should be a miniature version of the total test. That is, it should be composed of questions similar to the questions in the two forms to be equated.

Sample Size.

In testing programs in which large numbers of examinees are administered equivalent test forms, the statistical benefits of large samples (i.e., small standard errors) are realized. Furthermore, robustness to the violation of statistical assumptions is generally greater with large samples.

However, despite the fact that equating is generally conducted on large samples, the need for equating does not become insignificant when sample sizes are small. The effect on equating when small numbers of examinees complete equivalent forms has been insufficiently explored in the literature. The equating studies reported typically are conducted on several thousand examinees.

Two of the few studies that have examined the effects of linear equating on small sample sizes were conducted by Kolen (1985) and Parshall, Du Bose, and Kromrey (1992).

Although the literature in the area is sparse, the need for test equating does not become insignificant in testing programs in which the sample size is small (Parshall et al., 1992).

Indices of Equating Error

In order to determine the accuracy of equating, some type of evaluative index must be used. The accuracy of equating is defined as the statistical bias in equating resulting from the difference between the mean equated score computed from samples and the population value of the equated score.

Many evaluative indices have been mentioned in the research literature however, clear agreement on the most appropriate evaluative index to use in determining the accuracy of linear equating has not been established (Parshall et al., 1992).

In an effort to investigate equating accuracy, Budescu (1985) proposed a model that assessed the relationship between the length of the anchor test and the efficiency of the equating process. After examining the derivation of the equating equations associated with Angoff Model IV, Budescu noted that the correlation coefficients between the anchor test and the test forms were key components in the estimation of the equating parameters. Using the reliability of the total test, the correlations between the anchor test and the test forms, and the anchor test length, Budescu developed an index that indicated how much more efficient the equating procedure may become by increasing the length of the anchor test. Budescu indicated that the anchor test correlation is dependent upon the reliability of the total test and the relative length of its components.

The findings from this study suggested that the magnitude of the correlation coefficient between the anchor test and the unique components of each form is the single most important factor in determining the efficiency of the equating process.

Ideally the correlations between the anchor test and the total tests should be unity; however, this is seldom the case in practice. In an attempt to suggest the most appropriate linear equating design to use when unity is not reached, Woodruff (1989) analyzed three linear equating designs. Using internal and external anchors Woodruff indicated that the higher the correlations between the anchor test and the total tests the more accurate the equating adjustment. He also suggested that the correlation coefficient between the anchor test and the total test be used as a measure in selecting the most appropriate equating design. That is, some equating designs are more sensitive to a lack of content balance between the anchor test and the total test (Woodruff, 1989).

Budescu (1985) and Woodruff (1989) used the magnitude of the correlation coefficient between the anchor test and the total tests as an index of equating accuracy. Klein and Jarjoura (1985) used the root-mean-squared-error (RMSE) and the mean equating error (bias) that contributed to the RMSE as an index to evaluate the importance of using anchor items that were representative of the total test. The RMSE was given as the weighted standard deviations of the equated scores:

$$RMSE = \frac{\sqrt{\sum_i n_i (X_i - X'_i)^2}}{\sum_i n_i}$$ where n is the number of people who obtained score i, X_i is the i-th raw score, and X'_i is the equated score of X_i.. Bias was defined as the difference between the mean of the raw scores and the mean of the equated scores: $Bias = \overline{X_i} - \overline{X'_i}$, where $\overline{X_i}$ is the mean of the raw scores and $\overline{X'_i}$ is the mean of the equated scores.

On the other hand, Kolen (1985) derived large sample standard errors for Angoff Model IV with and without the normality assumption as an evaluative index of equating error that is due to examinee sampling. The standard error of equating was studied under two methods, the delta method (computer simulation) and the real data method (Efron's bootstrap). A computer simulation was conducted to model two forms of a professional certification test (nonsymmetric simulation) and two forms of an achievement test (symmetric simulation). Sample sizes were 100 and 250 examinees per form, respectively. The results of the simulation indicated that standard errors computed without the normality assumption are more accurate than the standard errors based on the normality assumption.

For the real data, Efron's bootstrap method of calculating the standard errors was used. Data were two forms of a 125-item multiple-choice professional certification examination with 30 common items. The forms were administered to 773 and 795 examinees, respectively. According to the author, under the normality assumption the standard errors are larger at the higher score points and smaller at the lower score points than those standard errors derived without the normality assumption. In general, the standard errors are smallest near the mean and increase for scores further away from the mean.

Also Kolen (1985) indicated that the standard errors for the bootstrap method without the normality assumption were nearly identical to the delta method without the normality assumption.

Parshall et al., (1992) used several of the indices mentioned above to evaluate the accuracy of equating using Angoff Model IV with small sample sizes. Parshall et al., (1992) used the correlation coefficient, the standard error of equating, and equating bias. The standard error was defined as the standard deviations of the obtained equated scores in the bootstrap sample. The formula for the standard error is

$$SE(\Theta_i) = \sqrt{\frac{\sum_j (\Theta_{ij} - \Theta_{i.})^2}{n-1}}$$

where $SE(\Theta_i)$ is the standard error for equated score i, Θ_{ij} is the obtained equated score i in sample j, and Θ_{i} is the mean equated score i over 1000 samples.

Statistical bias in equating was defined as the difference between the mean of the equated scores in the bootstrap samples and their corresponding population mean. Statistical bias was given as, $Bias(\Theta_i) = \Theta_{i.} - \Theta_i$ where $Bias(\Theta_i)$ is the statistical bias for equated score i, $\Theta_{i.}$ is the mean equated score i over 1000 samples, and Θ_i is the population equated score i.

Two parallel forms for each of five teacher certification examinations were used in this study. Sample sizes of 15, 25, 50 and 100 were examined. Employing a Monte Carlo design, one thousand samples of each size were drawn with replacement for each certification test. The authors indicated that the standard error of equating increased as sample size decreased, and equating bias was essentially insignificant.

The standard error index presented by Klein and Jarjoura (1985), and Parshall et al., (1992) are different versions of the equation for the standard deviation of equated scores. Klein and Jarjoura (1985) used the weighted standard deviations of the equated scores and Parshall et al., (1992) used an unweighted version of the same formula. In the Parshall et al.,(1992) study the standard errors were computed at each score point.

Kolen and Whitney (1982) and Jaeger (1980) suggested the use of indices that were quite different from the studies presented above. Kolen and Whitney (1982) in a comparison of four equating procedures, used a cross-validation statistic as an evaluative index. Twelve forms of the Test of General Educational Development (GED) were equated by four different equating methods. Pairs of test forms were administered to examinees in counterbalanced order. Approximately 200 examinees were used to equate the eleven forms of the GED to a base form. Using an independent equivalent group of examinees, scores from the cross-validation sample were converted to the base form score scale. The cross-validation statistic was computed as "the mean-squared difference (over examinees in the anchor form distribution) between anchor form integer scores and converted scores on the other form with identical percentile ranks in their respective cross-validation distributions" (Kolen & Whitney, p. 284). The formula for the cross-validation statistic, referred to as a percentile comparison index is $C = \frac{\sum_i (X_i - Y'_i)^2}{nk}$, where X_i is the i-th order observed score on the anchor form, Y'_i is the equated score on the cross-validation distribution of equated scores that has the same percentile rank as X_i, n is the number of observed scores, and k is the number of items on the anchor form. The authors concluded that the cross-validation procedure was effective in determining the relative accuracy of the equating methods studied even though the sample size was small.

Jaeger (1980) examined five statistical indices for their usefulness in selecting a test equating method. The indices were: (a) the similarity of two cumulative score distributions, (b) the shape of the raw score to scaled score transformation, (c) the consistency of linear and equipercentile equating results, (d) the similarity of the item difficulty distributions, and (e) the similarity of item discrimination distributions. The data for this study were gathered from the administration of a college aptitude test. Eight forms of the test were administered over a three year period. Five of the eight forms were used in equating. The sample size for the five equatings ranged from 5000 to 6000. According to Jaeger (1980) four of the five indices (the similarity of the two cumulative score distributions, the raw score to scaled score transformation, the consistency of linear and equipercentile equating results, and the similarity of the item difficulty distributions), distinguished between linear equating methods that were and were not adequate. Of the five indices, the similarity of item difficulty distributions seemed to be the most useful evaluative index.

Method

A Monte Carlo study was conducted to examine equating stability and statistical bias in a single and double linkage plan in small samples. Small random samples of size 25, 50 and 100 were drawn with replacement from archival test data files that represented Form B, Form N, and Form C pseudo-populations. Test data from two teacher certification subject area tests were used: Art Education(K-12), and Hearing Impaired (K-12). One thousand bootstrap samples were drawn with replacement for each pair of test forms, and sample size per subject area examination.

Descriptive statistics, and the correlations between the anchor test and the total test coefficients for each test form are presented in Tables 1 and 2, respectively. The test lengths range from 98 items to 110 items. The number of anchor items on each test is approximately 30% or more of the total test. The correlations presented in Table 2 are moderately high to high. According Budescu (1985) the correlation coefficients between the anchor test and the total test is an essential component in estimating equating parameter estimates.

Insert Table 1 and 2 about here

Using Angoff Model IV non-equivalent linear equating model, a direct link, an indirect link, and the average of the two links (direct and indirect) equating equations were computed for each pair of samples at each sample size, per subject area examination.

The stability of the equating linkage plans was evaluated by calculating the bootstrap standard errors of equating. A measure of statistical bias was used to evaluate the accuracy in the equating equations. The standard errors of equating are defined as the variability in equated scores resulting from sampling. Statistical bias in equating is defined as the difference between the mean equated scores computed following resampling and the population equated score.

The formula for the standard error is

$$SE(\Theta_i) = \sqrt{\frac{\sum_j (\Theta_{ij} - \Theta_{i.})^2}{n-1}}$$

where $SE(\Theta_i)$ is the standard error for equated score i, Θ_{ij} is the obtained equated score i in sample j, and $\Theta_{i.}$ is the mean equated score i over 1000 samples. Statistical bias in equating was defined as the difference between the mean of the equated scores in the bootstrap samples and their corresponding population mean. The corresponding population mean for this study is the Base-to-Current direct link.

Statistical bias was given as, $Bias(\Theta_i) = \Theta_{i.} - \Theta_i$ where $Bias(\Theta_i)$ is the statistical bias for equated score i, $\Theta_{i.}$ is the mean equated score i over 1000 samples, and Θ_i is the population equated score i.

Results

Equating Stability

The obtained estimates of the standard errors of equating for the direct link, the indirect link, and the averaged links for the Art and Hearing Impaired examinations are presented in Figures 2 through 7. These figures are graphical representations of the standard errors of equating (an index of equating stability) at all possible raw score points. Obvious in all the standard error figures is that the standard error of equating for each linkage plan is: (a) smallest at the mean, and increases

Insert Figures 2 - 7 about here

as a function of the deviation of scores away from the mean, and (b) equating stability decreases as sample size decreases. These results are consistent with results reported by Parshall et al, (1992). In viewing the stability of each linkage plan, it is noted that the indirect link is the least stable of all the links across sample sizes and examinations. For the Art examination the direct link and the averaged links behaved similarly across sample sizes. At raw score points below the mean the standard errors were slightly smaller for the direct link. On the other hand, at raw score points above the mean the averaged links evidenced smaller standard errors. For the Hearing impaired examination the averaged link provided the smallest standard error across all score points and across all sample sizes examined.

Statistical Bias in Equating

Graphs of the statistical bias in equating for the Art and Hearing Impaired examinations are presented in Figures 8 through 13. These figures are graphical illustrations of statistical

Insert Figures 8 - 13 about here

bias at all possible raw score points. The most striking characteristic of these figures was the magnitude of the equating bias for the direct link. That is, equating bias is basically trivial across all raw score points, regardless of the sample size. For all linkage plans, as sample size increased statistical bias decreased. These findings corresponds to findings presented by Parshall et al, (1992) in their study on small sample equating. Statistical bias in equating for the indirect link and the averaged links were quite large, relative to that observed for the direct link, with the indirect link showing the most bias in equating. Moreover, the pattern of the bias in the indirect and averaged links was consistent across test forms and sample sizes. Specifically, the equating was biased in a positive direction for low scores (i.e., below the mean) and negatively biased for high scores.

Discussion

In examining the standard errors and bias, the findings from this study indicate that: (a) the direct linkage design is much more stable across raw score points than the indirect linkage design, (b) equating bias for the direct linkage design is trivial, and (c) equating bias is quite large for the indirect linkage design. An advantage in terms of standard errors was observed when averaging the direct and indirect links for the Hearing Impaired examination. Such a reduction in standard error was not seen on the Art examination. However, when the two links were averaged, a substantial increase in equating bias was observed in both examinations.

As a result of the findings from this study (specifically the equating bias resulting from the indirect link), the authors recommend that the direct linkage design be used with small sample equating.

References

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), Educational Measurement (2nd ed.) (pp. 508-600). Washington, DC: American Council of Education.

Brennan, R. L., & Kolen, M. J. (1987). Some practical issues in equating. Applied Psychological Measurement, 11(3), 279-290.

Budescu, D. V. (1985). Efficiency of linear equating as a function of the length of the anchor test. Journal of Educational Measurement, 22(1),13-20.

Cope, R. T. (1987). How well do the Angoff Design V linear equating methods compare with the Tucker and Levine Methods? Applied Psychological Measurement, 11(2),143-149.

Jaeger, R. M. (1980, April). Some Exploratory Indices for Selection of a Test Equating Method. Paper presented at the annual meeting of the American Educational Research Association, Boston, MA.

Klein, L. W., & Jarjoura, D. (1987). The importance of content representation for common-item equating with non random groups. Journal of Educational Measurement, 22(3), 197-206.

Kolen, M. J., & Brennan, R. L. (1987). Linear equating models for the common-items nonequivalent-populations design. Applied Psychological Measurement, 11(3), 263-277.

Kolen, M. J., & Whitney, D. R. (1982). Comparison of four procedures for equating the tests of general educational development. Journal of Educational Measurement, 19(4), 279-293.

Kolen M. J. (1985). Standard errors of Tucker equating. Applied Psychological Measurement, 9(2), 209-223.

Parshall, C. G., Du Bose, P., & Kromrey, J. D. (1992, April). Common item linear equating in small samples of examinees: An empirical comparison of sample size effect. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R. L. Linn (Ed.), Educational Measurement (3rd ed.) (pp. 221-262). New York: National Council on Measurement in Education and American Council on Education.

Woodruff, D. J. (1989). A comparison of three linear equating methods for the common-item nonequivalent-populations design. Applied Psychological Measurement, 13(3), 257-261.

Table 1

Summary Descriptive Statistics for the Population of Teacher Certification Test Data Files

							Equating Links								
							B-N			B-C			N-C		
Exam Number	Content Area	Test Form	N of Examinees	Total Items	Raw Score Mean	Raw Score Std Dev	Anchor Items	MN	SD	Anchor Items	MN	SD	Anchor Items	MN	SD
01	Art (K-12)	Form B	401	105	68.00	8.82	67	42.86	6.11	50	32.72	5.09	**	*****	****
		Form N	388	109	70.61	8.66	67	43.95	5.71	**	*****	****	36	22.90	3.65
		Form C	131	113	73.42	9.60	**	*****	****	50	32.11	4.61	36	23.62	3.90
28	Hearing Impaired	Form B	36	98	70.14	7.04	43	30.42	3.63	51	36.92	4.07	**	*****	****
		Form N	59	102	69.07	8.47	43	29.24	4.34	**	*****	****	30	21.02	3.16
		Form C	46	107	70.24	9.94	**	*****	****	51	34.54	5.27	30	20.70	3.22

Table 2

Correlation Coefficients Between The Anchor Test and the Total Test by Subject Area

Exam Number	Content Area	Equating Links					
		B-N		B-C		N-C	
01	Art Education	Form B	.93	Form B	.91	Form N	.77
		Form N	.91	Form C	.92	Form C	.89
		B-N		B-C		N-C	
20	Hearing Impaired	Form B	.86	Form B	.92	Form N	.79
		Form N	.88	Form C	.94	Form C	.78

Figure 2
Standard Errors of Equating, Form=Art, N=25

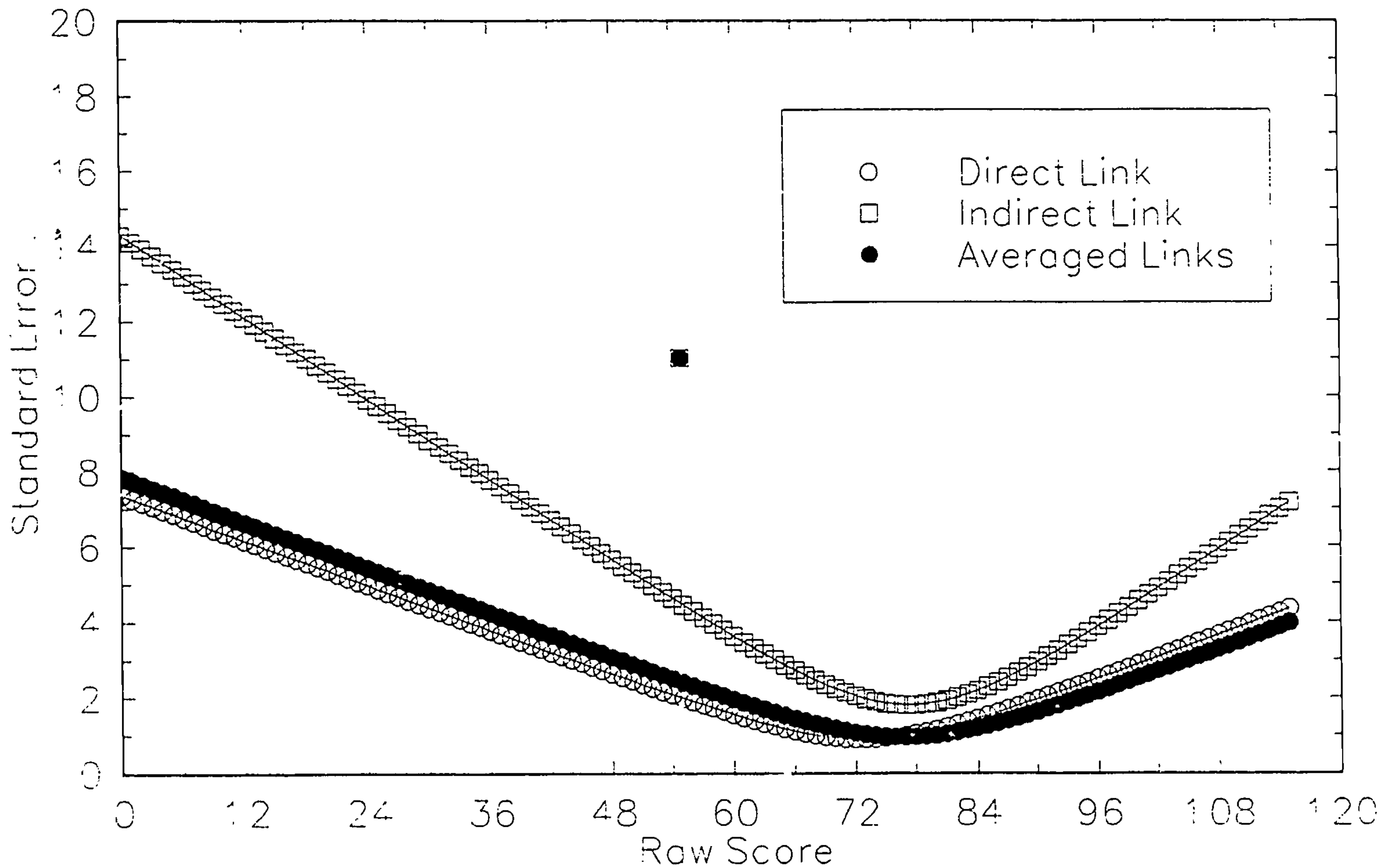

Figure 3

Standard Errors of Equating, Form=Art, N=50

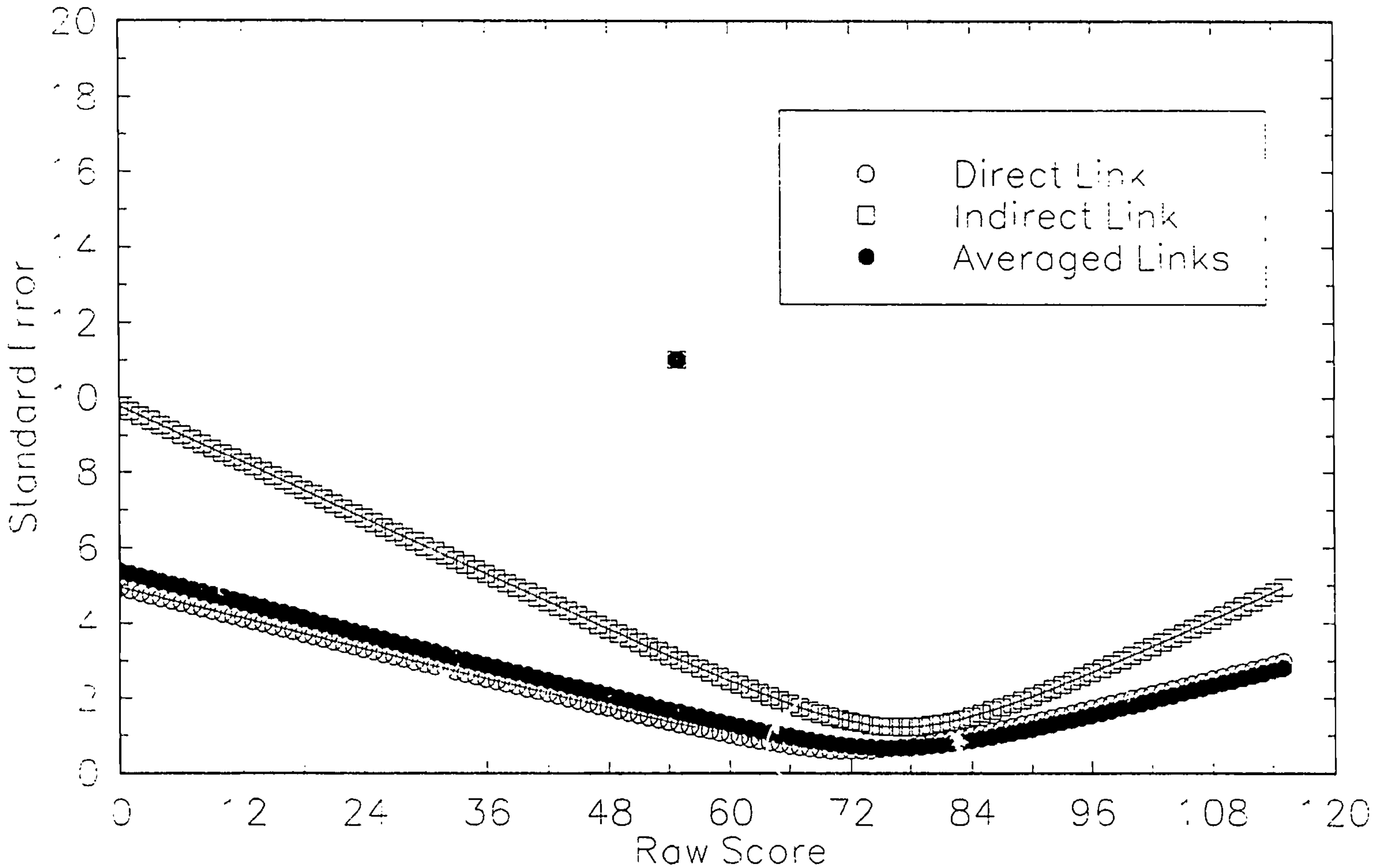

Figure 4
Standard Errors of Equating, Form=Art, N=100

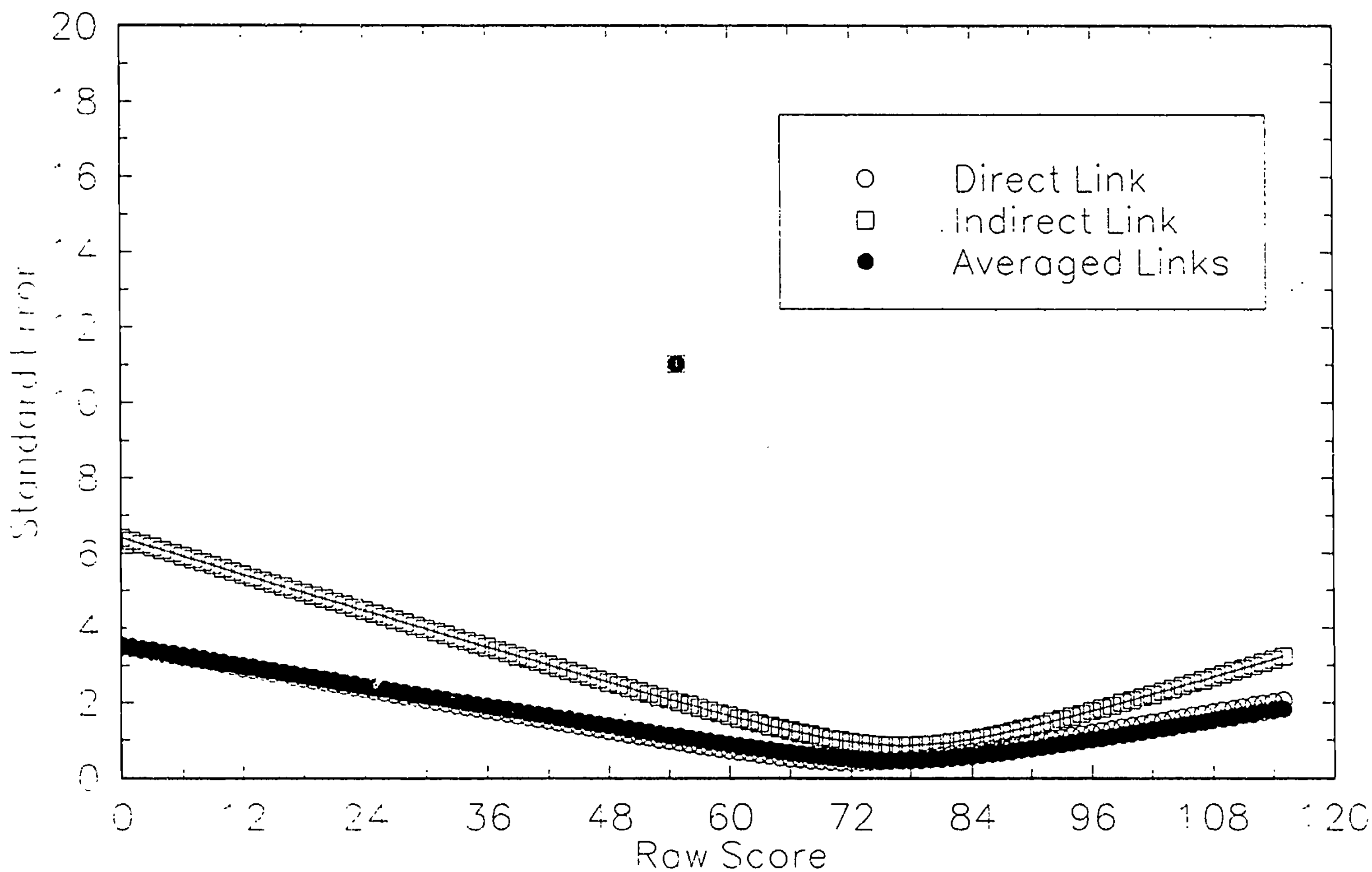

Figure 5
Standard Errors of Equating, Form=Hearing Impaired, N=25

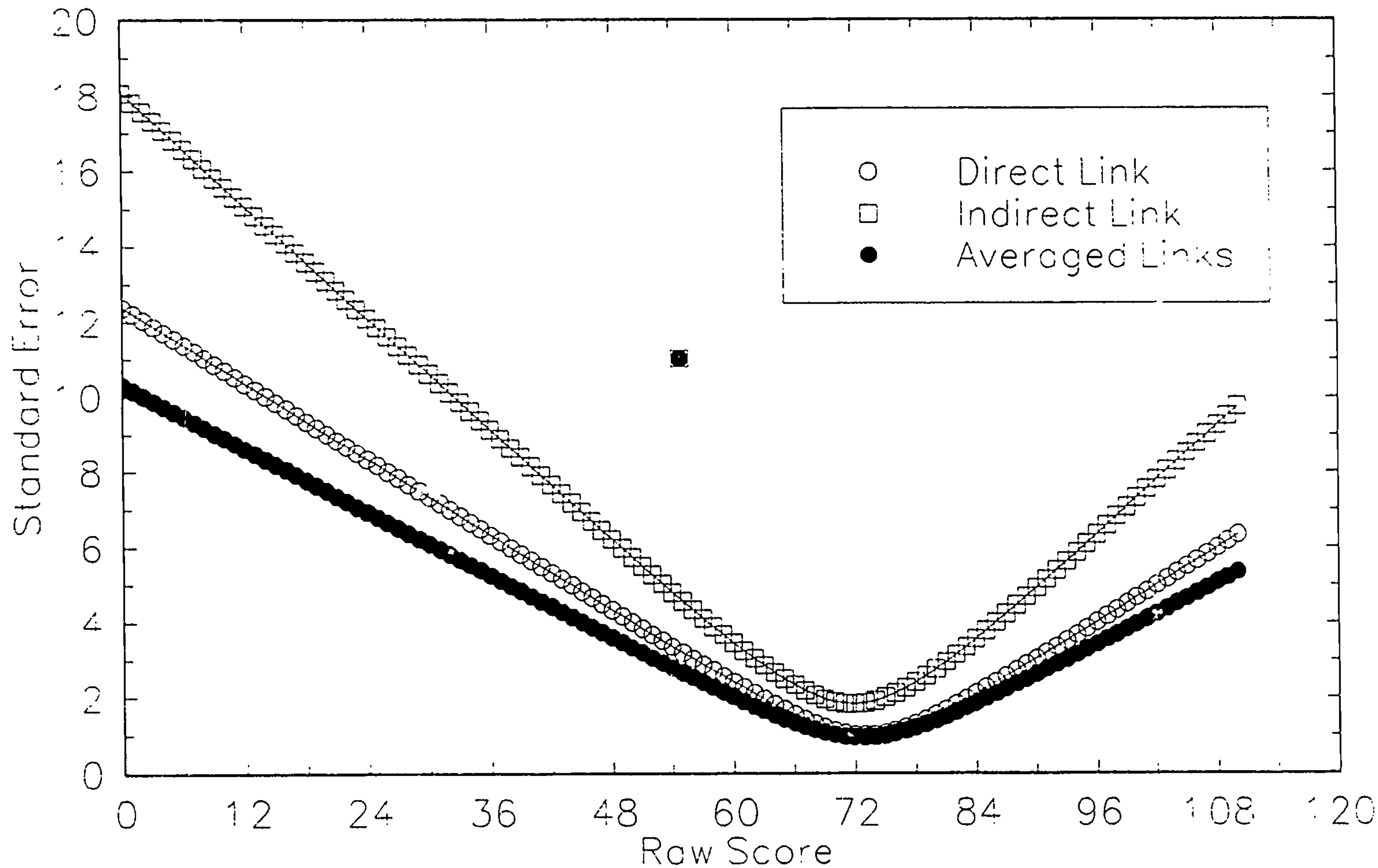

Figure 6

Standard Errors of Equating, Form=Hearing Impaired, N=50

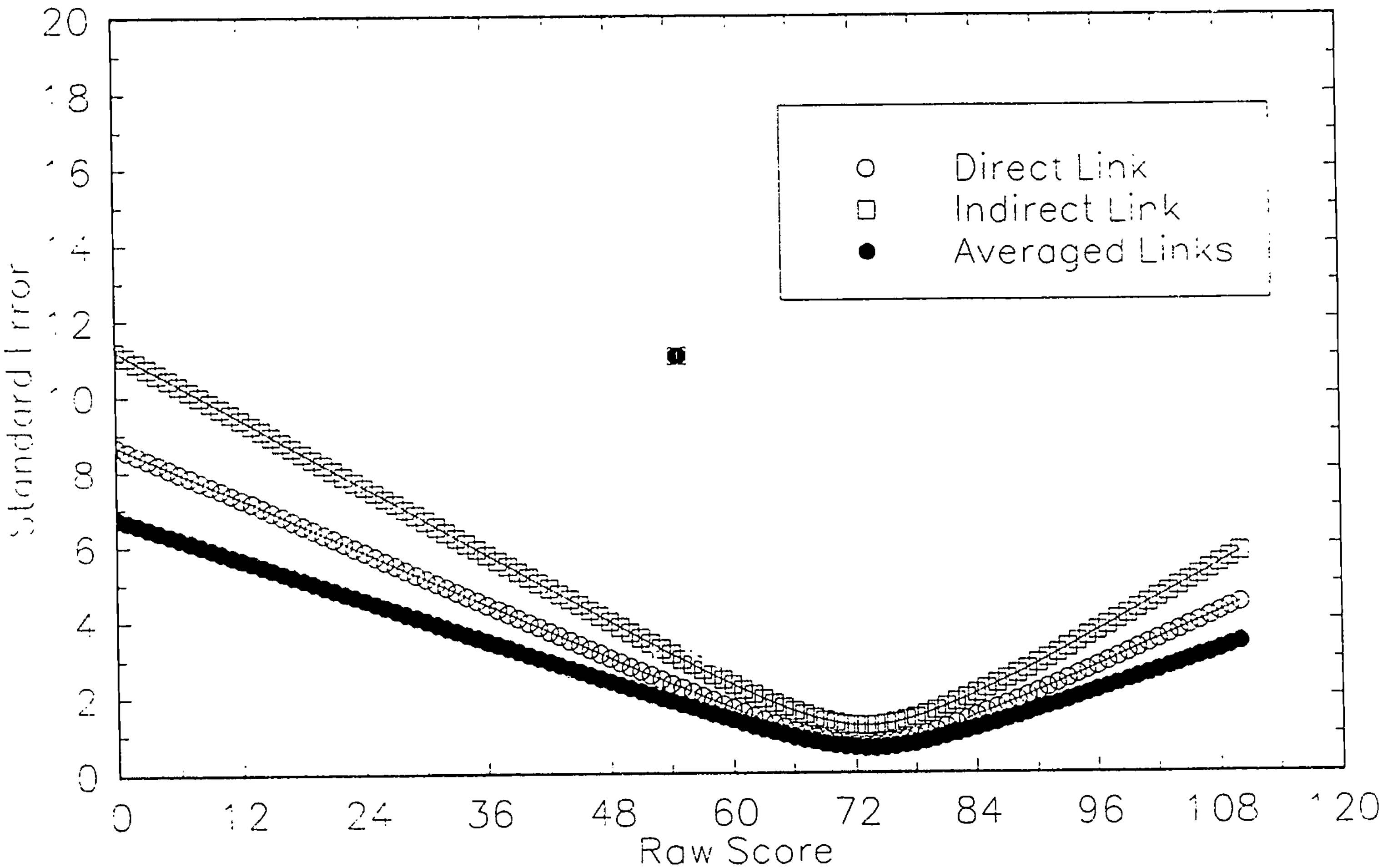

Figure 7

Standard Errors of Equating, Form=Hearing Impaired, N=100

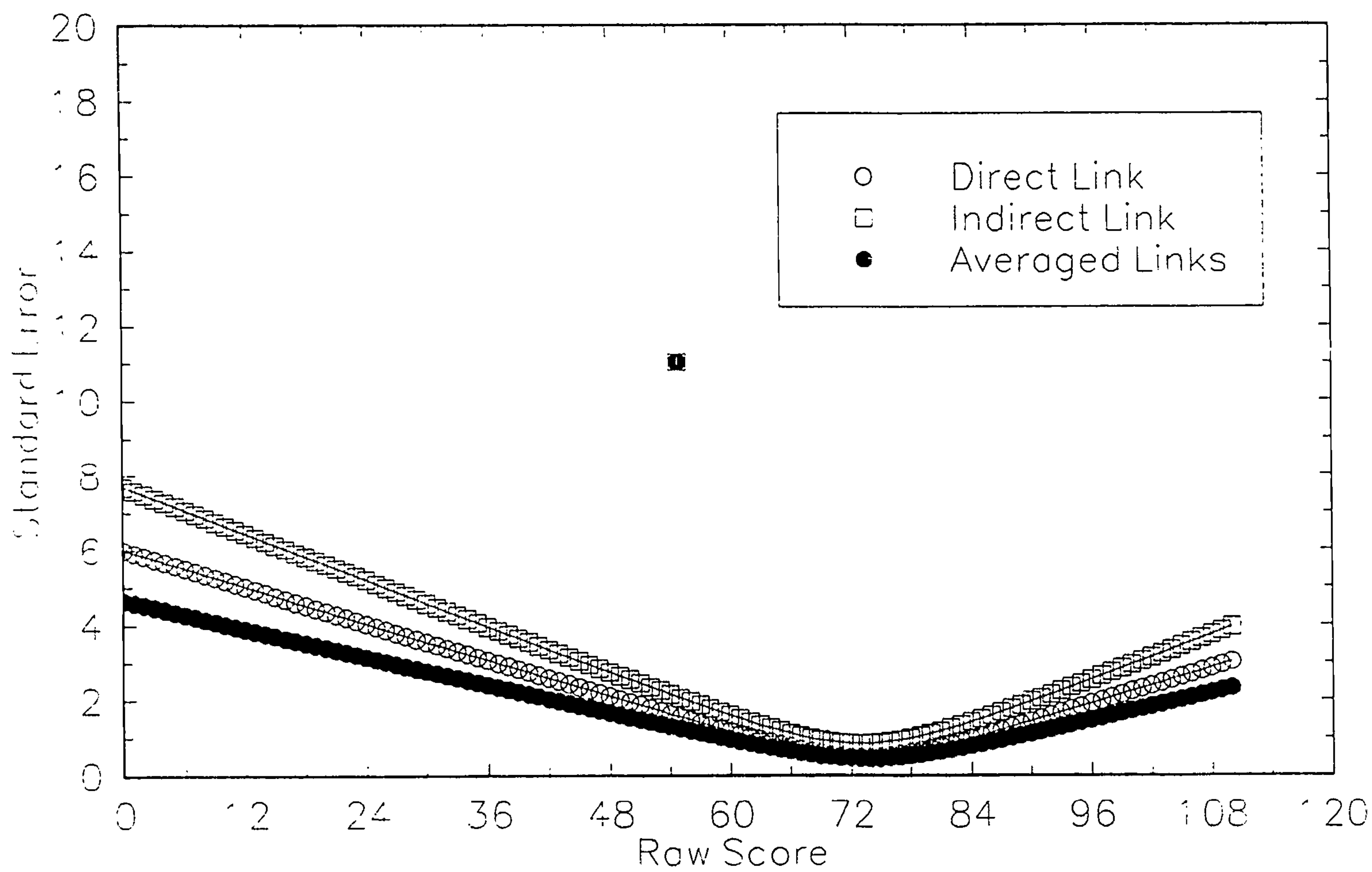

Figure 8
Equating Bias, Form=Art, N=25

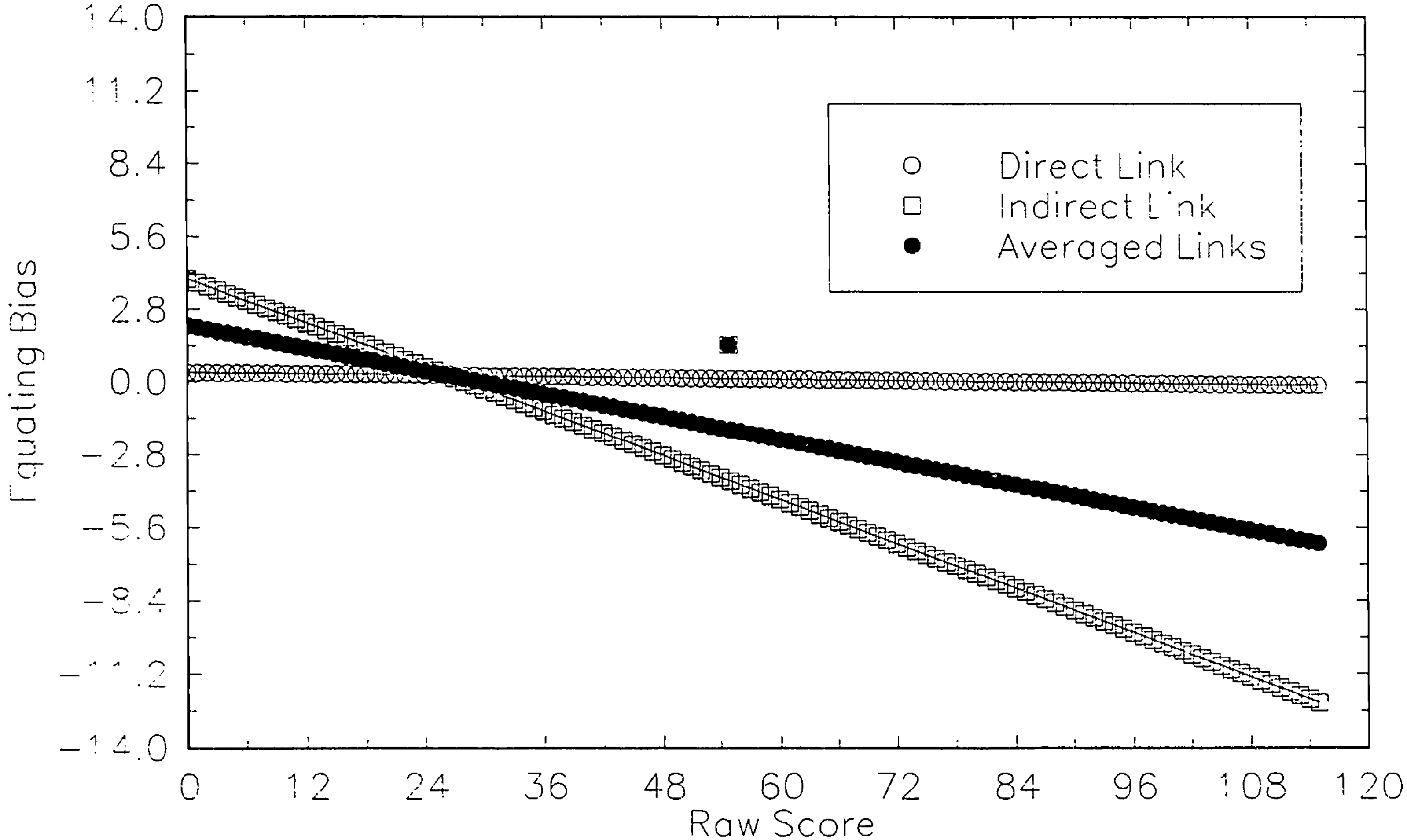

Figure 9

Equating Bias, Form=Art, N=50

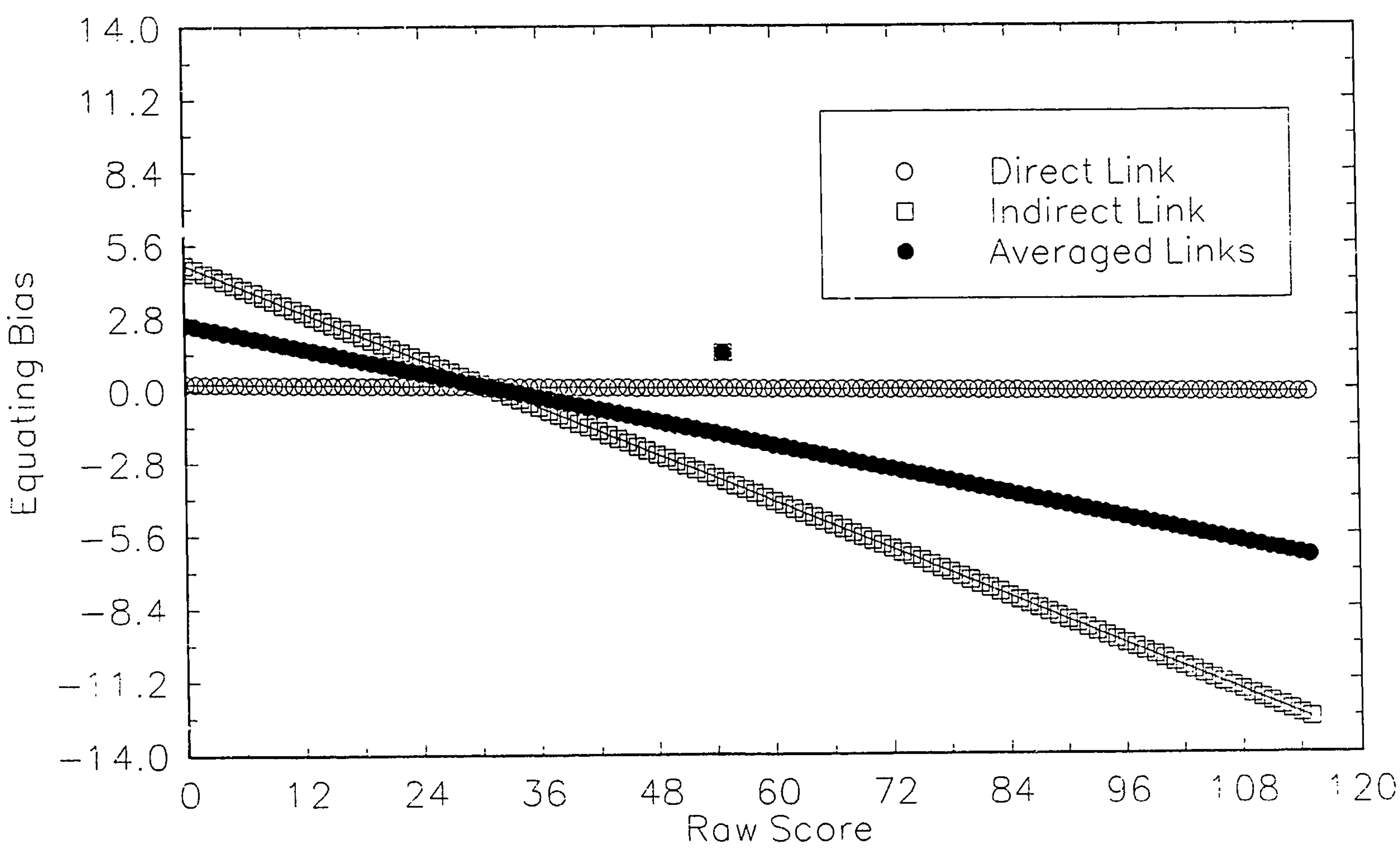

Figure 10
Equating Bias, Form=Art, N=100

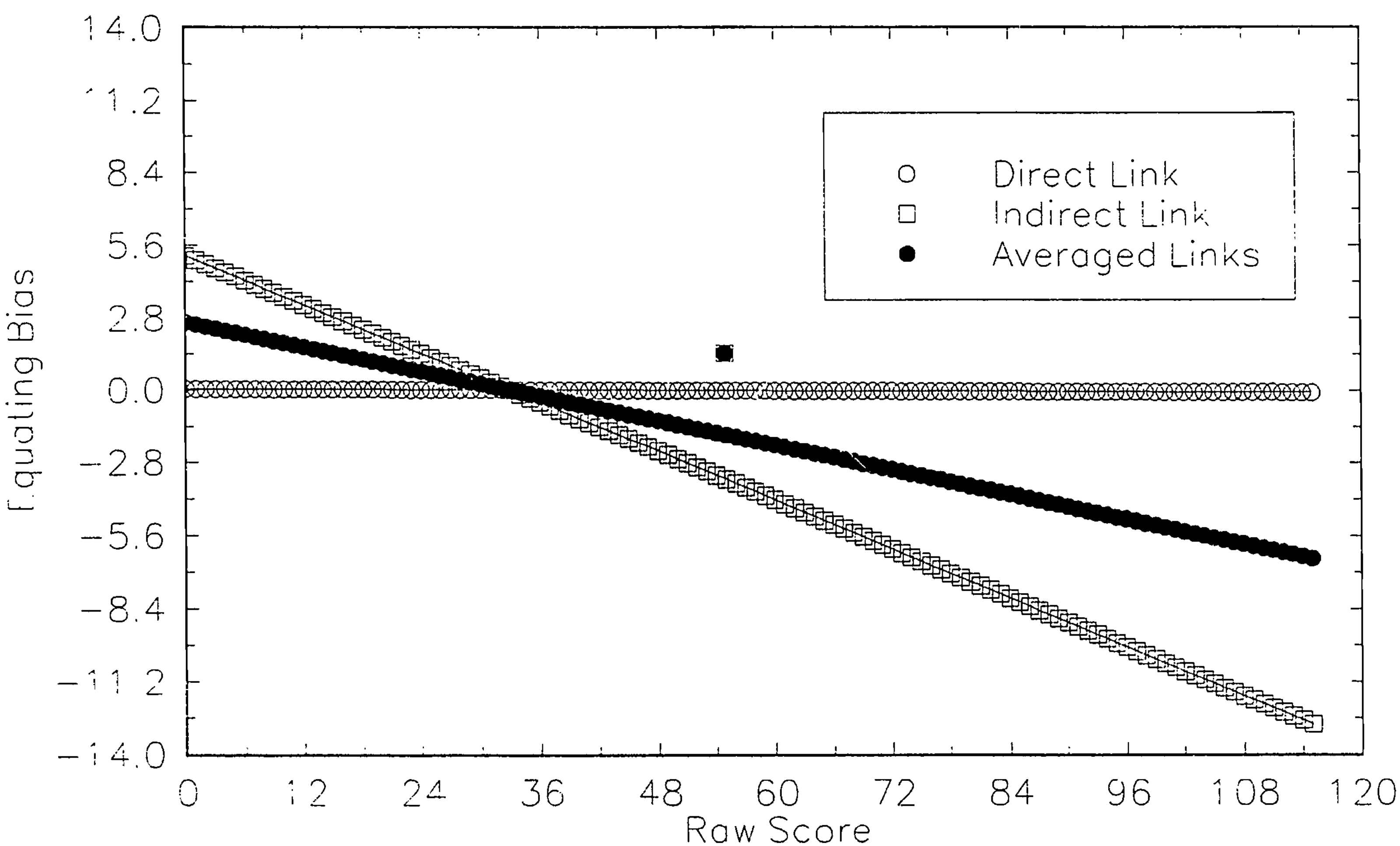

Figure 11

Equating Bias, Form=Hearing Impaired, N=25

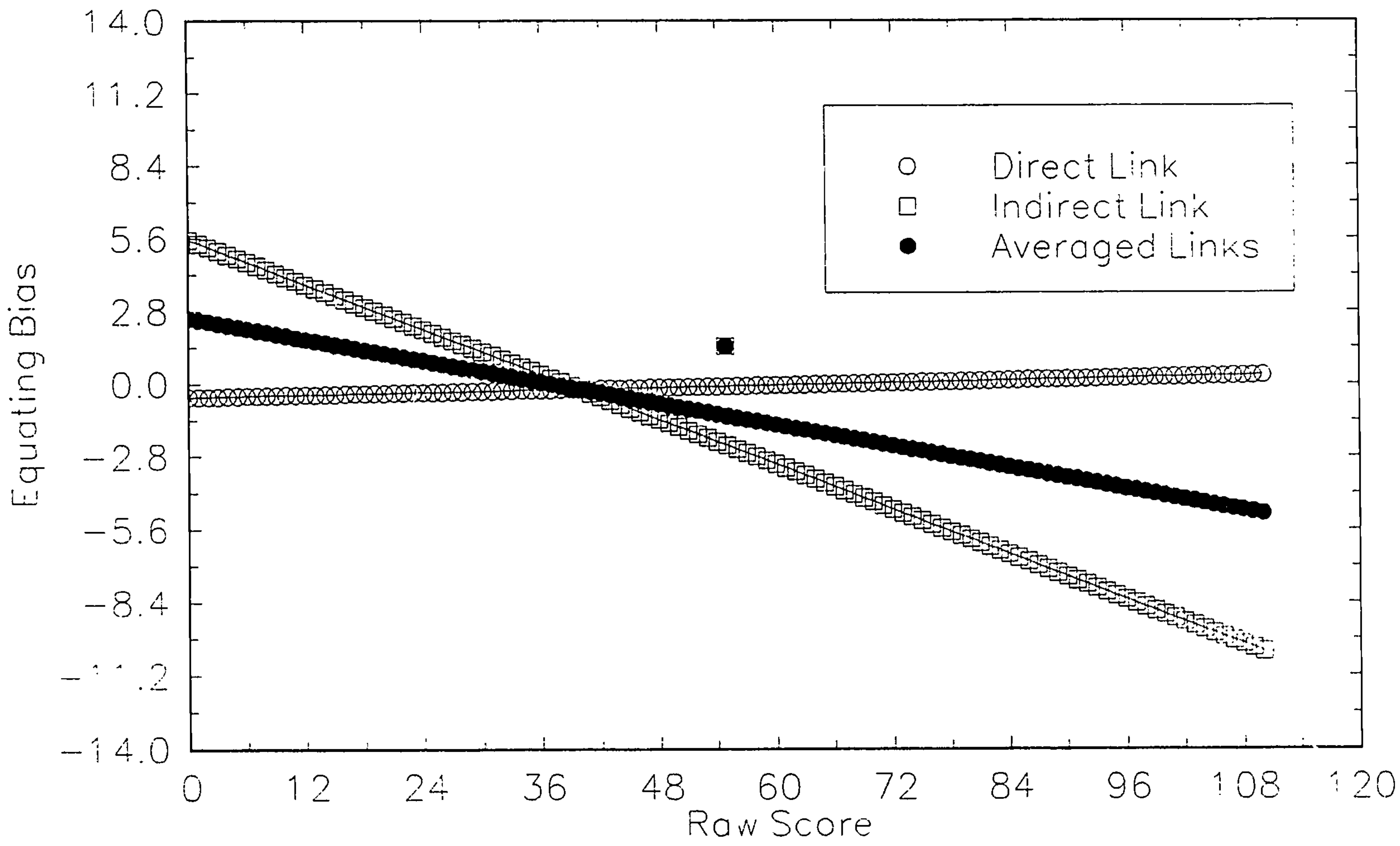

Figure 12

Equating Bias, Form=Hearing Impaired, N=50

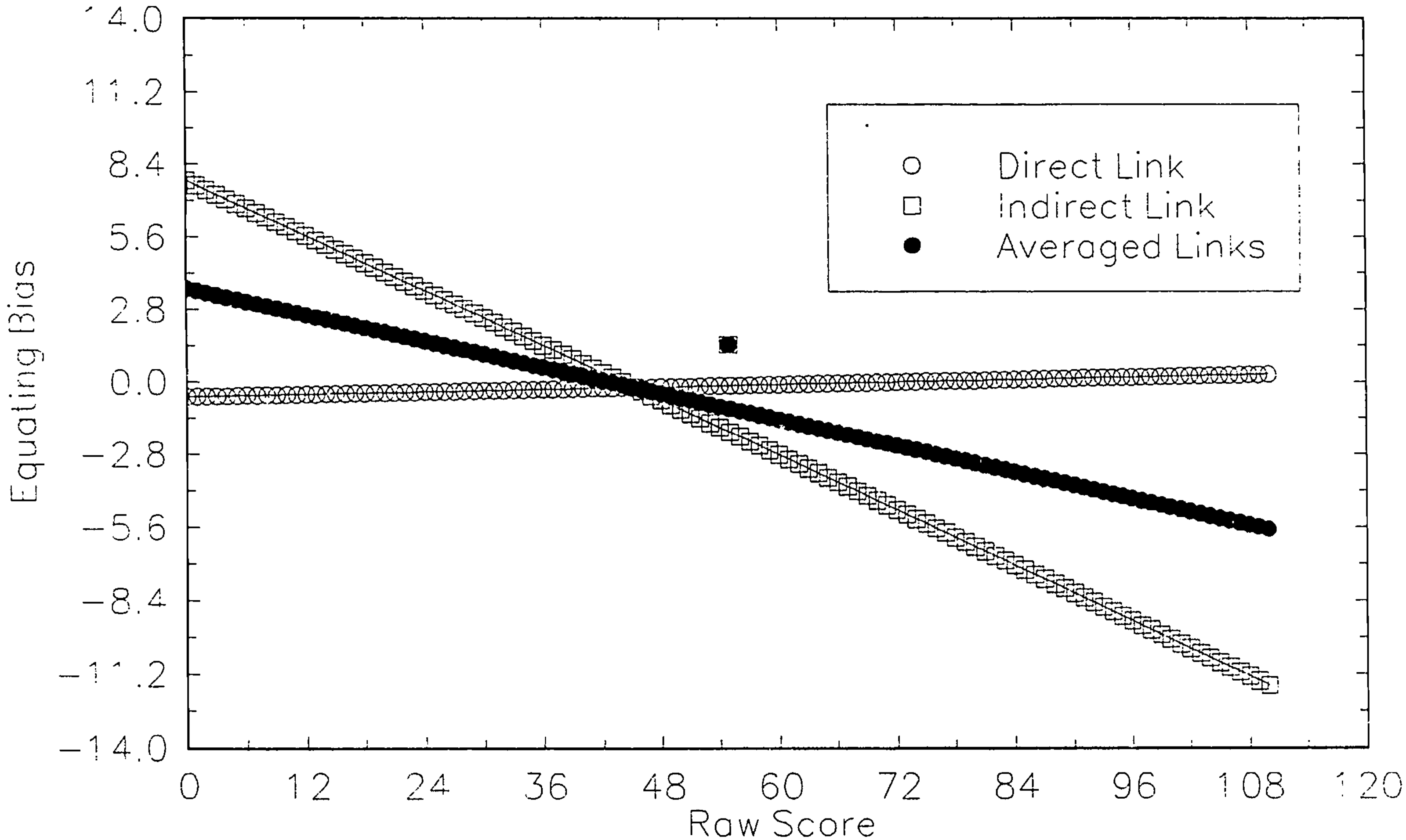

Figure 13

Equating Bias, Form=Hearing Impaired, N=100