9. CONTROL LINES FOR ITEM PLOTS

When tests intended to measure on a particular variable are used with different groups of persons or to measure persons under different conditions, it is necessary to determine the degree of stability the tests maintain over these occasions. The quantitative comparisons sought depend on the tests retaining the same quantitative definition of the variable throughout the occasions to be compared. In order to determine this, a method is required to evaluate the invariance of the common test item calibrations from group to group or time to time.

In order to evaluate the invariance of these calibrations we need to compare item calibrations to see whether quantitative comparisons of the measures obtained from these occasions are possible. To do this we need to compare the centered calibrations for the items common to the two occasions.

In this chapter we explain how to make such comparisons (1) by plotting the centered item calibration estimates from two different occasions against one another, (2) by analyzing the standardized differences of the item calibrations between the two occasions and (3) by evaluating the correlation between the pairs of estimates over the set of common items.

In order to be explicit, we follow our explanations with an example to help the reader work through each step in the process. In the previous chapter, Identifying Item Bias, we showed how to evaluate item bias through the use of item plots. That chapter concentrated on explaining the concepts involved and using the figures to illustrate the concepts. In this chapter we explain the techniques by which such plots are constructed and evaluated.

PLAN OF ACTION

1. Estimate the item calibrations for each of the two occasions and identify the set of items common to both occasions. These alternative calibrations may come from two different samples of persons or from the same sample of persons tested at two different times. Estimate the item calibrations with their respective estimation standard errors and fit statistics. Thus for each calibration occasion and for each item i we calculate the item difficulty estimate d_i, its standard error s_i, and the fit of the calibrating data to these estimates, v_i.

2. Center each set of common item calibrations on the same origin (using perhaps the mean difficulty of the common items in the most recent or most important test) so that their comparison becomes independent of any translation effects between the centers of the two calibrated tests.

   (If there is a translation, then that amount would have to be accounted for before person measures from the two occasions could be compared. See Wright and Stone, 1979, pp. 96-98 and 112-117. The best way to proceed, however, is to carry out a third calibration of all of the data from both previous calibrations pooled into one combined data matrix. Usually this combined data matrix, in which every item on either test defines a column of possible responses and every individual test administration in either sample defines a row, has some empty cells where that item was not administered to that person. The "missing" data is easily managed in a calibration program like BIGSTEPS, (Wright, 1996)).

3. Plot these paired and centered item calibrations d_{1i} and d_{2i} against one another for each common item. A common variable is demonstrated when the plotted item points, which should estimate a single common difficulty for each item, fit an identity line, e.g. fall within one or two standard errors of their identity line.

4. Construct statistical control lines around the identity line by computing standard units of error along lines which are perpendicular to the identity line and passing through the item points. (The error control lines can be constructed for one or two error units producing 68% and 95% quality control.)

   These control lines can be used to evaluate, at a glance, the overall stability of the item calibrations shown on the plot. If more item calibrations fall outside the control lines than are expected by the control choices of 68% or 95%, we are led to doubt the stability of the calibrations in this study and to investigate the particular items causing the visible lack of invariance. Even when only a few items fall outside the control lines, we examine the particulars of these items carefully to determine why this has occurred and what we might do to control these particular conditions which threaten the validity of measurements made with these items.

5. Calculate the standardized difference between the alternate estimates of the single common item difficulty:

$$z_{12i} = (d_{1i} - d_{2i}) / (s_{1i}^2 + s_{2i}^2)^{1/2} .$$

   This statistic has an expectation of zero and a variance of one when item stability holds. The pattern of these differences can be studied by plotting z_{12i} against $d_{.i} = (d_{1i} + d_{2i}) / 2$.

6. Correlate d_{1i} with d_{2i} over the $i = 1, L$ common items. This correlation r_{12} has a maximum value governed by the standard errors s_{1i} and s_{2i} and also the variance of the $d_{.i}$. This maximum correlation is:

$$R_{\max} = 1 - (SE^2 / SD^2) = 1 - [(L-1)/L] * [\sum_i^L (s_{1i}^2 + s_{2i}^2) / \sum_i^L (d_{1i} + d_{2i})^2]$$

$$\text{when } d_{1.} = d_{2.} = 0$$

$$SE^2 = \sum_i^L (s_{1i}^2 + s_{2i}^2) / 4L$$

$$SD^2 \sum^L (d_{1i} + d_{2i})^2 / 4(L-1)$$

Fisher's log transformation for linearizing correlations can be used to compare the observed correlation r_{12} with the maximum correlation $R_{\max}$ in order to test the hypothesis of item calibration stability.

$$t = \left[\frac{(L-3)^{1/2}}{2} \right] \log \left[\frac{(1+r_{12})(1-R_{\max})}{(1-r_{12})(1+R_{\max})} \right]$$

This statistic has expectation zero and variance one when item stability holds. It tests for the overall fit of these L items to the identity line which defines invariance.

AN EXAMPLE

These steps are illustrated in the following tables and figures. There is a first test form of 14 items calibrated on a sample of 34 persons. Then the variable was expanded by the development of 10 additional items making a second test form of 14 + 10 = 24 items which is given to a sample of 101 persons. The original 14 items remain common to both forms of the test. We evaluate the stability of the 14 items between these two test forms to determine whether the two item calibrations are statistically equivalent and so can be combined to define measures on a single common variable.

If this contention is supported by our analysis, then we can compare and pool the measures of the original 34 persons with the measures of the later 101 persons producing a sample of 135 persons measured on the same variable.

If, however, this contention is not supported, then we cannot compare or pool the original 34 measures with the subsequent 101 measures because we have found them to be measured on different variables. Then we are forced to review how these items are functioning in order to discover why the items are not working the way we intended.

1. Table 9.1 gives the item calibrations for each test form. The old and new item names for Forms 1 and 2 are given in Columns 1 and 5 with the old item calibrations for Form 1 listed in Column 2 and the new item calibrations for Form 2 listed in Column 6. The new item names for Form 2, given in Column 5, are shown with their old Form 1 item names in parentheses. These new item calibrations for the 14 original items are given again in Column 7.

   Observe that the center (mean) of the 14 Form 1 old item calibrations is at 0.0 (Column 2) and the center (mean) of the 24 Form 2 new item calibrations is also at 0.0 (Column 6). These zeros, however, are not equivalent, since the old zero defines the center of the old 14 items while the new zero defines the center of the new 24 items. In fact, the center (mean) of the new Form 2 calibrations for the 14 original items is now 0.4 on the new scale of Form 2 (Column 7). Because of this difference the calibrations of the original 14 items must be shifted by 0.4 (Column 3). This shift puts them on the same scale as the new 24 items and produces the adjusted values given in Column 4 which are the values that will be used to compare item stability between Forms 9.1 and 9.2.

2. The adjusted Form 1 (Column 4) and Form 2 (Column 7) calibrations of these 14 items are plotted in Figure 9.1. The plot shows that these items fall along the identity line rather well,

Table 9.1

Comparing the Calibrations of 14
Items Common to Two Test Forms

FIRST TEST FORM				SECOND TEST FORM		
(1) Old Item Name	(2) Old Item Calibration	(3) Shift Value	(4)* Adjusted Calibration	(5) New Item Name	(6) New Item Calibration	(7)** New Calibration (Original 14 Items)
				1	-6.0	
				2	-5.6	
1	-4.2	0.4	-3.8	3 (1)	-3.8	-3.8
2	-3.6	0.4	-3.2	4 (2)	-2.3	-2.3
3	-3.2	0.4	-2.8	5 (3)	-2.5	-2.5
				6	-4.0	
4	-3.6	0.4	-3.2	7 (4)	-2.3	-2.3
5	-2.2	0.4	-1.8	8 (5)	-1.8	-1.8
6	-3.2	0.4	-2.8	9 (6)	-1.8	-1.8
7	-1.5	0.4	-1.1	10 (7)	-0.8	-0.8
				11	0.1	
				12	-0.6	
				13	-0.3	
				14	-1.3	
				15	-0.5	
8	0.8	0.4	1.2	16 (8)	2.2	2.2
9	2.1	0.4	2.5	17 (9)	1.6	1.6
10	1.9	0.4	2.3	18 (10)	2.2	2.2
11	3.2	0.4	3.6	19 (11)	3.1	3.1
12	4.6	0.4	5.0	20 (12)	3.6	3.6
13	4.6	0.4	5.0	21 (13)	3.6	3.6
14	4.6	0.4	5.0	22 (14)	4.7	4.7
				23	6.5	
				24	6.0	
Column Mean	0.0		0.4		0.0	0.4
SD	3.4		3.4		3.4	2.8

* (4) = (2) + (3)

The comparison will be made between (4) and (7).

** (6) = (7)

but, as yet, we have no way to evaluate how much these item plots could deviate from the exact identity line before we would be forced to decide that the differences are too much. To accomplish this evaluation, we construct quality control lines. These lines guide our study of the plot to help us to make useful decisions.

3. Figure 9.2 lays out a simple way to construct these control lines. The standard unit of difference error parallel to either axis for item i is:

$$S_{12i} = (s_{1i}^2 + s_{2i}^2)^{1/2}$$

The notes appending Figure 9.2 give the details for determining the coordinates (X and Y) for a machine plot of the control lines. See Table 9.3 for application to our data. Entering these values in a plotting program can produce smoothed quality control lines.

Table 9.2 shows how to do a simple hand plot of the control lines. This is used with our sample data and shown in Figure 9.3.

A unit of error equivalent to S_{12i} but perpendicular to the 45 degree identity line is:

$$T_{12i} = \left[\left(s_{1i}^2 + s_{2i}^2\right)/2\right]^{1/2} = S_{12i}/\sqrt{2}$$

One of these T error units perpendicular to the identity line, through the $(d_{1i},\ d_{2i})$ item plot and extended in each direction from the identity line yields a pair of 68% control lines. Two of these T error units perpendicular to the identity line yields a pair of 95% control lines.

Table 9.2 gives the standard errors s_{1i} and s_{2i} (Columns 6 and 7) for the 14 common items connecting Forms 1 and 2.

We calculate T_{12i} for each of the 14 items and plot these locations in Figure 9.3 at two standard error units above and below the identity line. These points can be connected and smoothed to provide the quality control lines needed to evaluate the item plots.

4. Figure 9.3 shows that the plots of the 14 items of Forms 1 and 2 are all well within two standard errors of the identity line. It also shows that the hand and constructed methods of drawing in control lines lead to identical results. We conclude that these 14 items fall along the identity line, given their standard errors. Our variable extension is successful according to this sample data.

5. We can also evaluate the standardized item calibration differences between the Form 1 and Form 2 item calibration estimates for these 14 items by using:

$$Z_{21i} = (d_{2i} - d_{1i})/(s_{1i}^2 + s_{2i}^2)^{1/2}.$$

These standardized differences are expected to have a mean of zero and a variance of one. The standardized differences of the 14 items are given in Column 9 of Table 9.2. Trends can be evaluated by plotting these Z_{21i} against $d.i$ for each item.

Figure 9.4 is this plot. We observe that all of the remaining items are well within 1.0. All

Figure 9.1

Plot of common item calibrations: Form 1 versus Form 2.

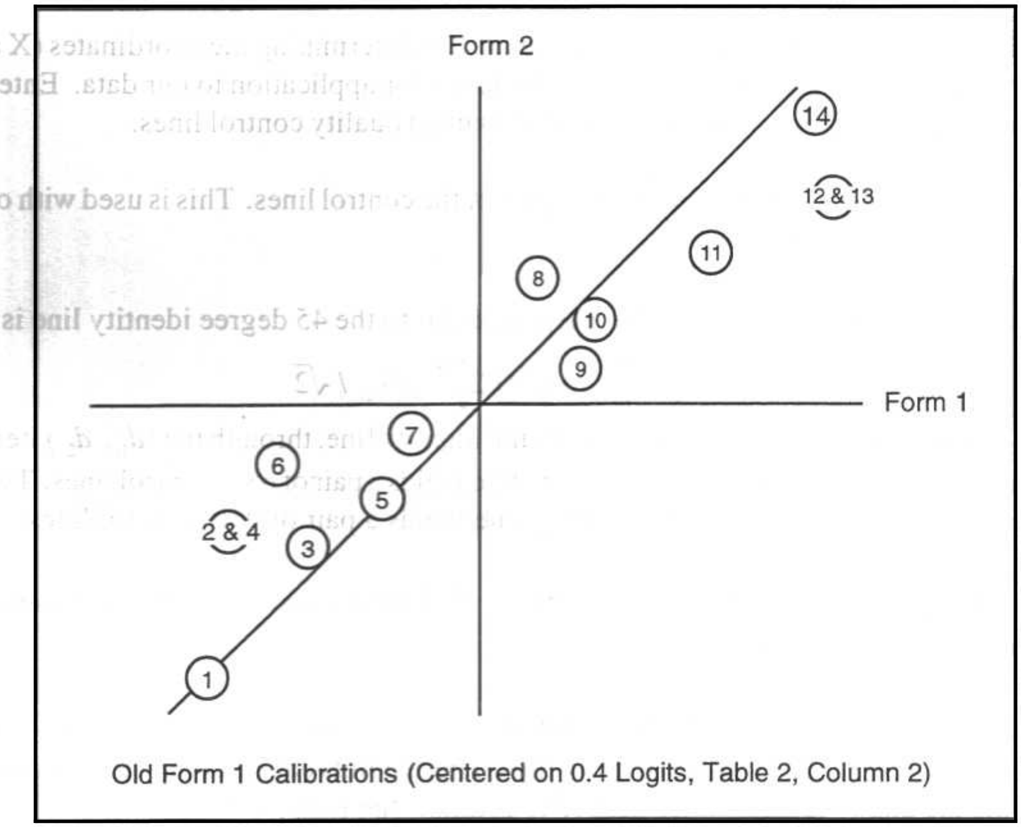
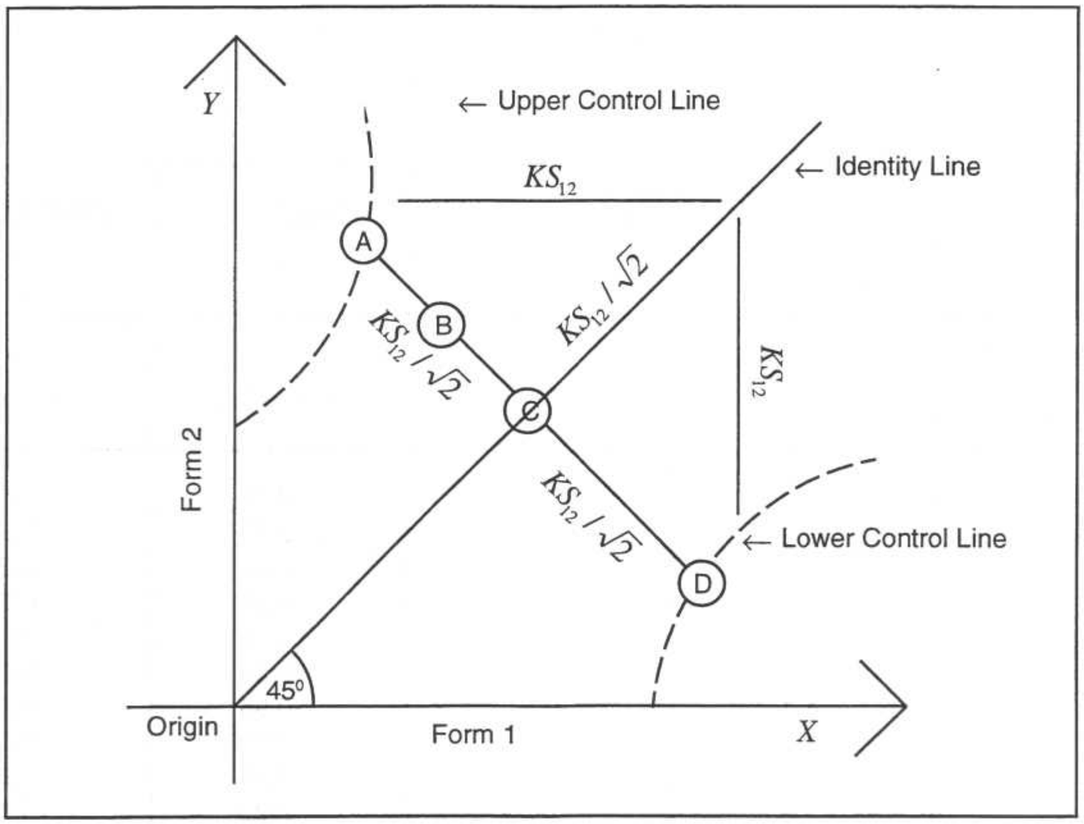


Figure 9.2

How to construct control lines.



Upper Control Line:

Position A: $X = d - KS_{12}/2 = (d_1 + d_2 - KS)/2;\ Y = d + KS_{12}/2 = (d_1 + d_2 + KS_{12})/2$

Item Plot:

Position B: $X = d_1;\ Y = d_2$

Identity Line:

Position C: $X = (d_1 + d_2)/2 = d;\ Y = (d_1 + d_2)/2 = d$

Lower Control Line;

Position D: $X = d + KS_{12}/2 = (d_1 + d_2 + KS_{12})/2;\ Y = d - KS_{12}/2 = (d_1 + d_2 - KS_{12})/2$

See Table 9.3 and Figure 9.3 for an example.

K = number of standard error units chosen to set the confidence level control of the lines; e.g., $K = 1$ produces 68% confidence and $K = 2$ produces 95% confidence.

$S_{12} = \sqrt{S_1^2 + S_2^2}$ the standard error of the difference $(d_1 - d_2)$

S_1 = the standard error of d_1
S_2 = the standard error of d_2
d = $(d_1 + d_2)/2$

Table 9.2.

Item Calibrations, Standard Errors
and Standardized Differences Z

	CALIBRATION		AVERAGE $d_{\cdot i}$	DIFFERENCE
(1) Old Item Name	(2)* d_{1i}	(3)** d_{2i}	(4) $(d_{1i}+d_{2i})/2$	(5) $(d_{1i}-d_{2i})$
1	-3.8	-3.8	-3.80	0.0
2	-3.2	-2.3	-2.75	-0.9
3	-2.8	-2.5	-2.65	-0.3
4	-3.2	-2.3	-2.75	-0.9
5	-1.8	-1.8	-1.80	0.0
6	-2.8	-1.8	-2.30	-1.0
7	-1.1	-0.8	-0.95	-0.3
8	1.2	2.2	1.70	-1.0
9	2.5	1.6	2.05	0.9
10	2.3	2.2	2.25	0.1
11	3.6	3.1	3.35	0.5
12	5.0	3.6	4.30	1.4
13	5.0	3.6	4.30	1.4
14	5.0	4.7	4.85	0.3
MEAN***	0.4	0.4		
S.D.	3.4	2.8		

* Column 4 from Table 1

** Column 7 from Table 1

*** Items have been centered at the common mean for Form 2 of 0.4. This separates the analysis of the calibration differences $(d_{1i}-d_{2i})$ from any overall difference in test form difficulty.

Table 9.2. (Continued).

	STANDARD ERROR		STANDARD ERROR OF DIFFERENCE	STANDARDIZED DIFFERENCE	ERROR UNIT
Old Item Name	(6) S_{1i}	(7) S_{2i}	(8) S_{12i}	(9) Z_{21i}	(10) $T_{12i} = S_{12i}/\sqrt{2}$
1	0.8	1.0	1.28	0.00	0.91
2	0.7	0.7	0.99	0.91	0.70
3	0.7	0.7	0.99	0.30	0.70
4	0.7	0.7	0.99	0.91	0.70
5	0.5	0.6	0.78	0.00	0.74
6	0.7	0.6	0.92	0.88	0.81
7	0.5	0.6	0.78	0.29	0.74
8	0.4	0.8	0.89	0.92	0.77
9	0.5	0.6	0.78	-0.86	0.74
10	0.5	0.8	0.94	-0.09	0.81
11	0.7	0.8	1.06	-0.41	0.86
12	1.1	1.0	1.49	-0.97	1.03
13	1.1	1.0	1.49	-0.97	1.03
14	1.1	1.2	1.63	-0.20	1.05
			MEAN S.D.	+0.02 0.71	

$S_{12i} = (S_{1i}^2 + S_{2i}^2)^{1/2}$

$Z_{21i} = (d_{2i} - d_{1i})/S_{12i}$

Column (9) = (5)/(8)

The Quick Hand Method for Adding Control Lines

To draw 95% control lines by hand use the approximation (Wright and Stone, 1979, p. 95) for an error allowance perpendicular to the identity line:

$$2T_{12i} = [(S_{1i}^2 + S_{2i}^2)/2]^{1/2} \approx (S_{1i} + S_{2i}).$$

Mark off a piece of graph paper to match the plotting axes and then slide this special ruler along the identity line marking off the perpendicular distances $(S_{1i} + S_{2i})$ in each direction away from the identity line as each item point (d_{1i}, d_{2i}) is encountered. This is done in Figure 9.3 where the results are marked as small circles.

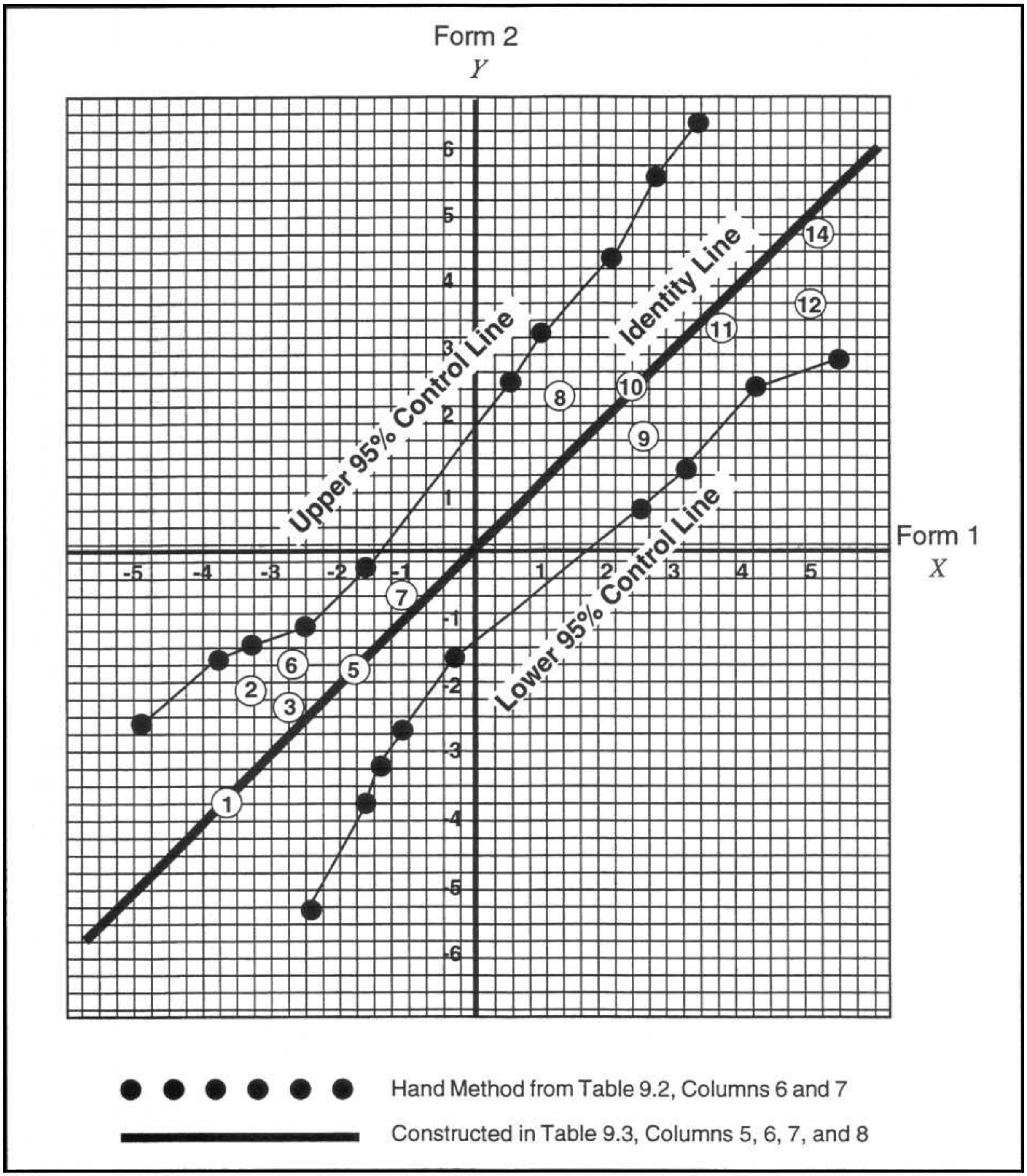
of the values are within the 68% control lines.

The correlation over $i = 1, 14$ of the calibrations d_{1i} and d_{2i} can also be determined. A limit for this coefficient is R_{max}. In our example $R_{\text{max}} = 0.98$ and the correlation for the observed item calibrations is also 0.98. Since $R_{\text{max}} = 0.98$ is the same as $r_{12} = 0.98$, we see that the correspondence between the item calibration estimates computed from the Form 1 and Form 2 samples is as good as can be expected. This correlation, when evaluated for its statistical deviation from the intended equating of Form 1 and Form 2 using Fisher's log transformation, produces a $T \approx 0.0$. We retain the hypothesis of no statistical difference between these 14 pairs of item calibrations and hence of the stability of these items and the variable they define over the two occurrences. As a result we can pool and compare the 34 and 101 person measures.

Our example has illustrated the steps for evaluating the stability of item calibrations. In our example we confirmed the invariance of our item calibrations. If confirmation were not achieved, we could not undertake any quantitative comparisons of the measures from the two occasions and it would be necessary to determine why particular items failed to support our intention to equate Form 1 and Form 2 and to compare the measures they produced. Changes might be made in these items or new items constructed and the equating process repeated with a new sample. Even when changes do not appear necessary it is prudent to monitor item calibration stability continually as new samples occur in order to verify that conditions have not changed.

Figure 9.3

Plot of item calibrations: Form 1 versus Form 2 with 95% control lines using hand method and constructed method of Table 9.3.



● ● ● ● ● ● Hand Method from Table 9.2, Columns 6 and 7

▬ Constructed in Table 9.3, Columns 5, 6, 7, and 8

Table 9.3

Example Data for Constructing 95% (K=2) Control Lines

Old Item Name	Item Plot *Figure 2 (B)*		Identity Line *Figure 2 (C)*	Standard Error	95% Upper Control Line *Figure 2 (A)*		95% Lower Control Line *Figure 2 (D)*	
	d_{1i} *(X)* (1)	d_{2i} *(Y)* (2)	d_i *(X,Y)* (3)	S_{12i} (4)	$d-S_{12i}$ *(X)* (5)	$d+S_{12i}$ *(Y)* (6)	$d+S_{12i}$ *(X)* (7)	$d-S_{12i}$ *(Y)* (8)
1	-3.8	-3.8	-3.80	1.28	-5.08	-2.52	-2.52	-5.08
2	-3.2	-2.3	-2.75	0.99	-3.74	-1.76	-1.76	-3.74
3	-2.8	-2.5	-2.65	0.99	-3.64	-1.66	-1.66	-3.64
4	-3.2	-2.3	-2.75	0.99	-3.74	-1.76	-1.76	-3.74
5	-1.8	-1.8	-1.80	0.78	-2.58	-1.02	-1.02	-2.58
6	-2.8	-1.8	-2.30	0.92	-3.22	-1.38	-1.38	-3.22
7	-1.1	-0.8	-0.95	0.78	-1.73	-0.17	-0.17	-1.73
8	1.2	2.2	1.70	0.89	0.81	2.59	2.59	0.81
9	2.5	1.6	2.05	0.78	1.27	2.83	2.83	1.27
10	2.3	2.2	2.25	0.94	1.31	3.19	3.19	1.31
11	3.6	3.1	3.35	1.06	2.29	4.41	4.41	2.29
12	5.0	3.6	4.30	1.49	2.81	5.79	5.79	2.81
13	5.0	3.6	4.30	1.49	2.81	5.79	5.79	2.81
14	5.0	4.7	4.85	1.63	3.22	6.48	6.48	3.22

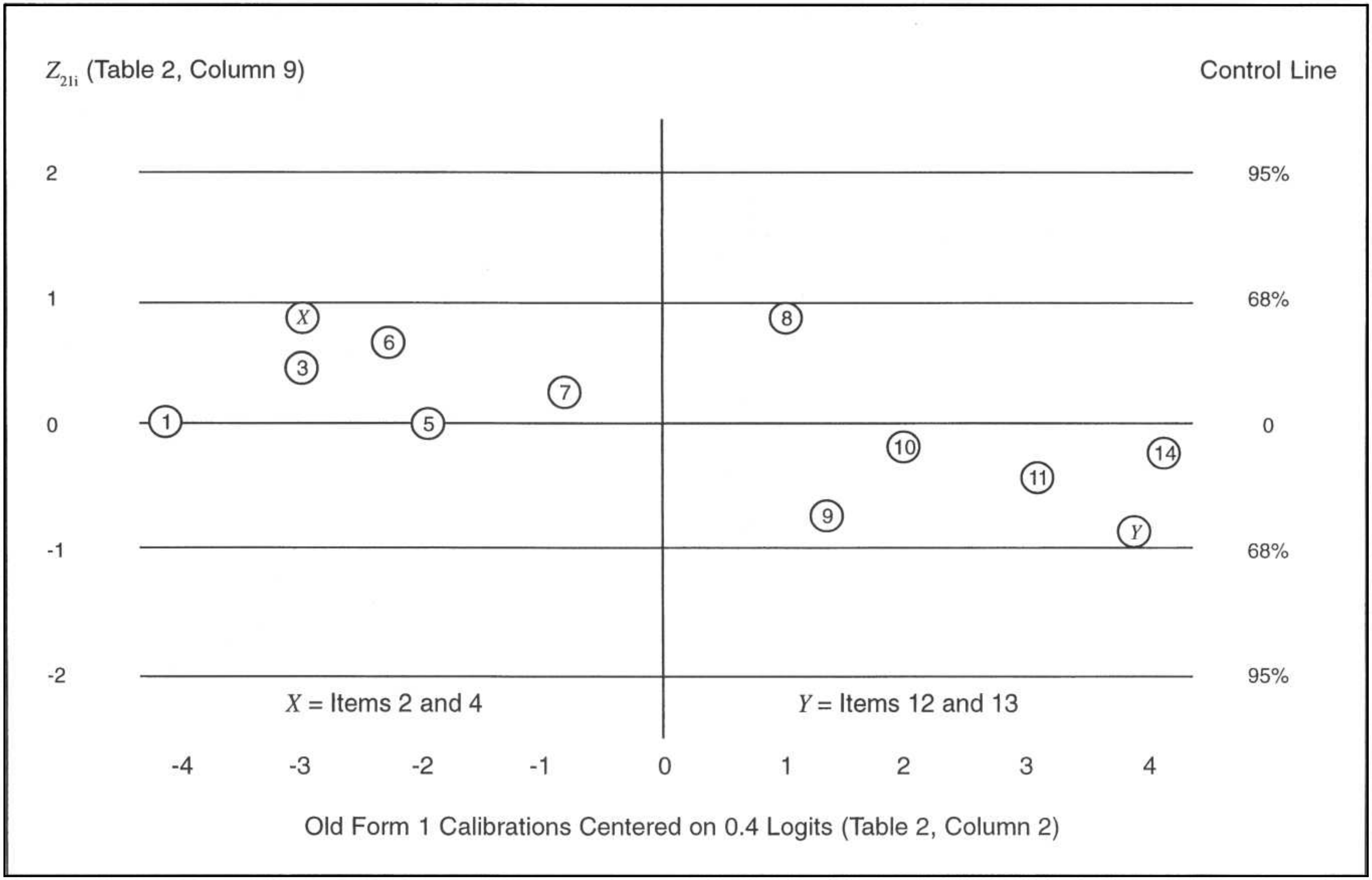
Columns 1 and 2 are from Table 2, Columns 2 and 3.

$d_I = (d_{1i} + d_{2i})/2$ (Table 2, Column 4)

$S_{12i} = (S_{1i}^2 + S_{2i}^2)^{1/2}$ (Table 2, Column 8)

Figure 9.4

Plot of item calibrations vs. standardized difference with 68% and 95% control lines.



MEASUREMENT ESSENTIALS

2nd Edition

BENJAMIN WRIGHT

MARK STONE

WIDE RANGE, INC.
Wilmington, Delaware