11. CONNECTING TESTS

In this chapter we describe the basic strategies for connecting tests intended to measure on the same variable so that the separate measures each test implies are expressed together on one single common scale. The process begins by understanding how to link two tests. Next we consider how to connect several tests and from there we proceed to plans for connecting all possible tests.

CONNECTING TESTS

The traditional method for connecting two tests is by equating the equal-percentile scores of a sample of persons who take both tests simultaneously. This process requires a large sample of persons with scores broadly enough distributed to assure an adequate representation of each score-to-percentile connection.

Rasch measurement enables a more economical and better controlled method for connecting tests and building item banks. Links of 10 to 20 common items are embedded in pairs of tests composed of otherwise different items. Each test is administered to its own sample of persons. No person need take more than one test. But all items in all tests can be subsequently connected through the network of common item links.

The traditional approach to equating two 60-item tests, say Test A and Test B, is to give both tests simultaneously to a sample of many, say 1200, persons as in Figure 11.1. The large sample is to assure the detailed representation of score percentiles necessary for successful percentile equating. Each person takes Test A and Test B, a total of 120 items.

In contrast, the Rasch approach can do the same job with each person taking only one test of 60 items. To accomplish this a third 60-item test, C, is made up of 30 items from each of the original tests A and B. Then each of these three tests is given to a sample of 400 persons as depicted in the lower half of Figure 11.1. Now each person takes only one test, but all 120 items are calibrated together through the two 30-item links connecting the three tests. The testing burden on each person is one half of that required by the equal percentile plan. But the equating of the tests is under far better control. In actual practice, the three samples can also be halved to 200 each without loss of control. This reduces the amount of data to one fourth of that required for the equal percentile equating.

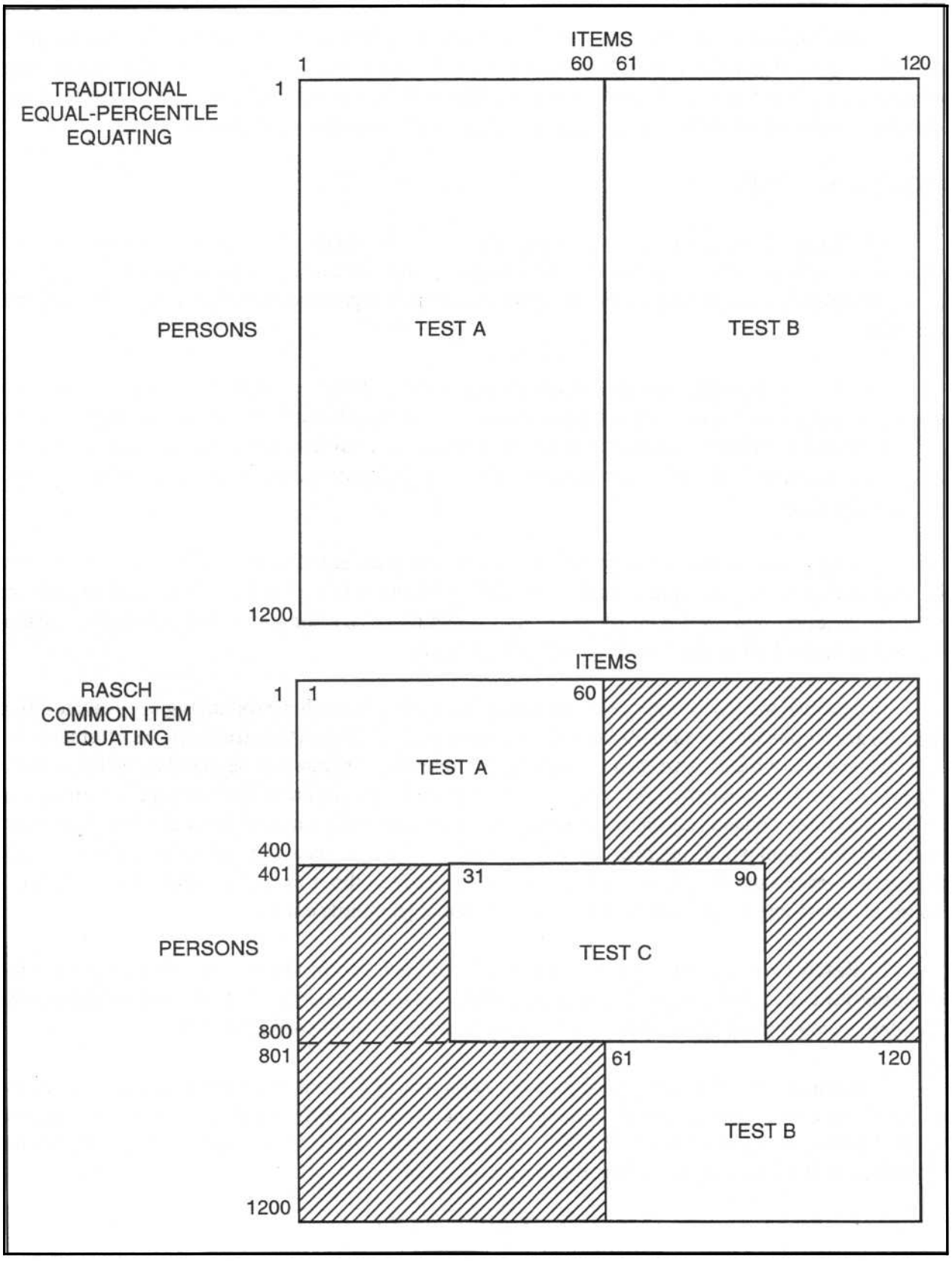
In Rasch equating, the separate calibrations of each test produce a pair of independent item difficulties for each linking item. The equating model asserts that each pair of estimates are statistically equivalent except for a single constant of translation common to all pairs in the link.

If two tests, A and B, are joined by a common link of K items and each test is given to its own sample of N persons, then d_{iA} and d_{iB} can represent the estimated difficulties of item i in each test with standard errors of approximately $2.5/N^{1/2}$ and the single constant necessary to translate all item difficulties in the calibration of Test B onto the scale of Test A is

$$G_{AB} = \sum_{i}^{K} (d_{iA} - d_{iB}) / K$$

Figure 11.1

Traditional and Rasch equating designs.



with standard error of approximately $3.5/(NK)^{1/2}$ logits.

In contrast to traditional equating, in which no quality control is available, the quality of this Rasch link can be evaluated by the fit statistic:

$$\sum_{i}^{K}(d_{iA}-d_{iB}-G_{AB})^2(N/12)[K/(K-1)]\sim X_K^2$$

which, when the two tests do fit together, will be distributed approximately chi-square with K degrees of freedom.

In addition, the individual fit of each item link can be evaluated by

$$(d_{iA}-d_{iB}-G_{AB})^2(N/12)[K/(K-1)]\sim X_1^2$$

which, when the performance of that item is consistent with the equating, will be approximately chi-square with one degree of freedom.

These simple fit statistics enable detailed, item by item control and remediation of test equations.

When using these chi-square statistics to judge link quality we keep in mind how they are affected by sample size. When N exceeds 500 these chi-squares can detect link flaws too small to make any noteworthy difference in G_{AB}, too small to matter. (When calibration samples are large, the root mean square misfit is more useful. This statistic can be used to estimate the logit increase in calibration error caused by link flaw.)

In deciding how to act on evaluations of link fit, we also keep in mind that random uncertainty in item difficulty of less than .3 logits has no discernible bearing on person measurement (Wright & Douglas, 1975, 35-39).

Because of the way sample size enters into the calculation of item difficultly and hence into the evaluation of link quality, we can deduce from these considerations that samples as small as 200 persons and links of as few as 10 good items will always be more than enough to supervise link validity at better than .3 logits. In practice we have found that we can construct useful and stable item banks with sample units as small as 50 persons.

THE COMMON LINK

The basic structure required to calibrate many items onto a single variable is the common item *link* in which one set of linking test items is shared by and so connects together two otherwise different tests. An easy and a hard test can be linked by a common set of intermediate items. These linking items are the "hard" items in the easy test but the "easy" items in the hard test (Figure 11.2).

With two or more test links we can build a *chain* of the kind shown in Figure 11.3.

The representation in Figure 11.3 can be conveyed equally well by the simpler scheme shown in Figure 11.4 which emphasizes the links and facilitates diagraming more complicated

Figure 11.2

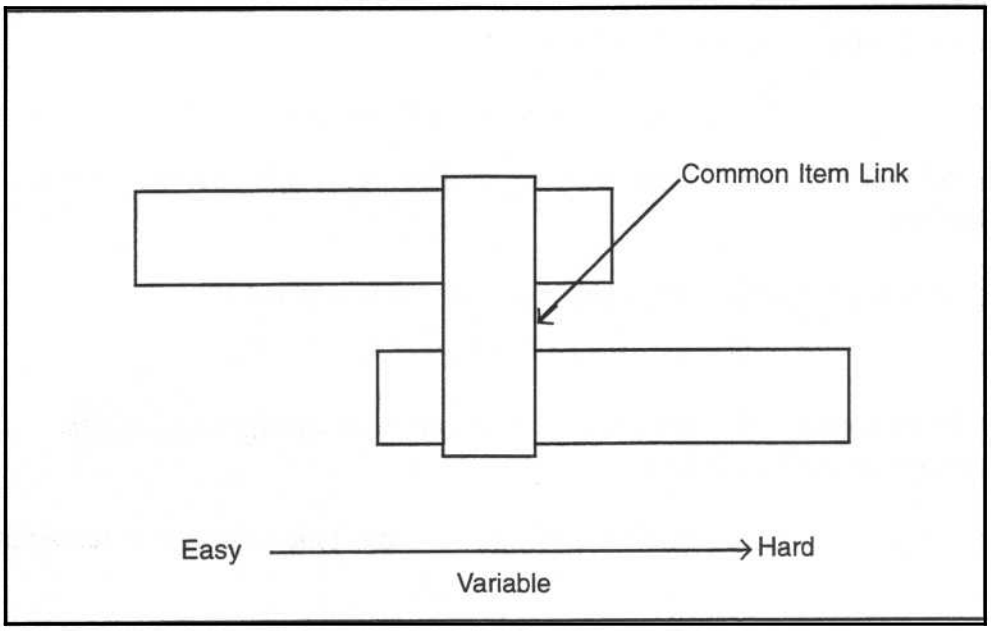
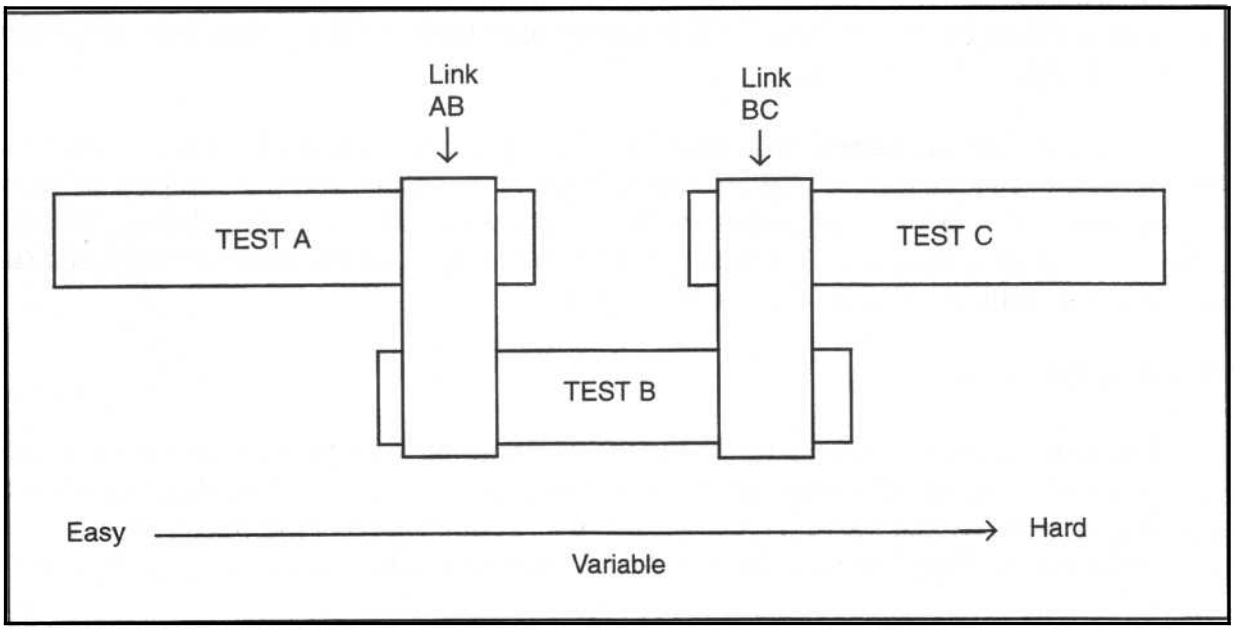


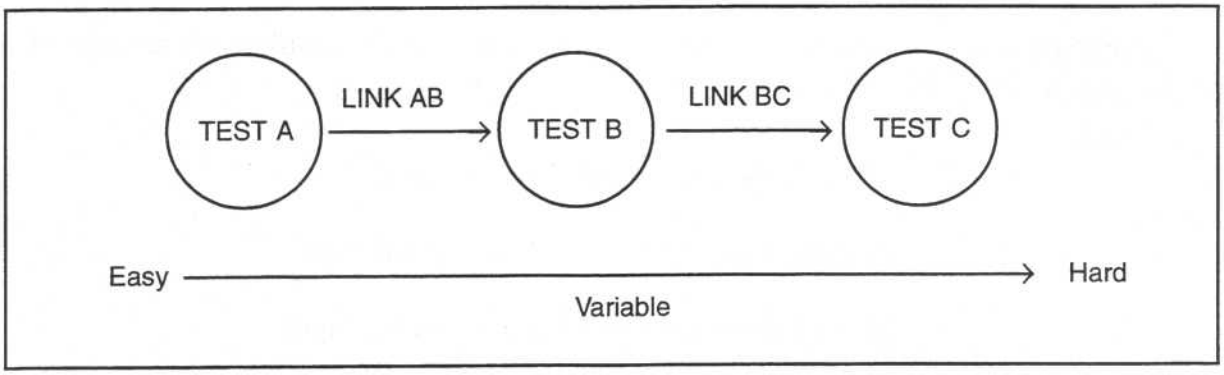
Figure 11.3



linking structures. Each circle indicates a test sufficiently narrow in range of item difficulties to be manageable by a suitably chosen sample of persons.

Figure 11.4

A chain with two links (simplified).

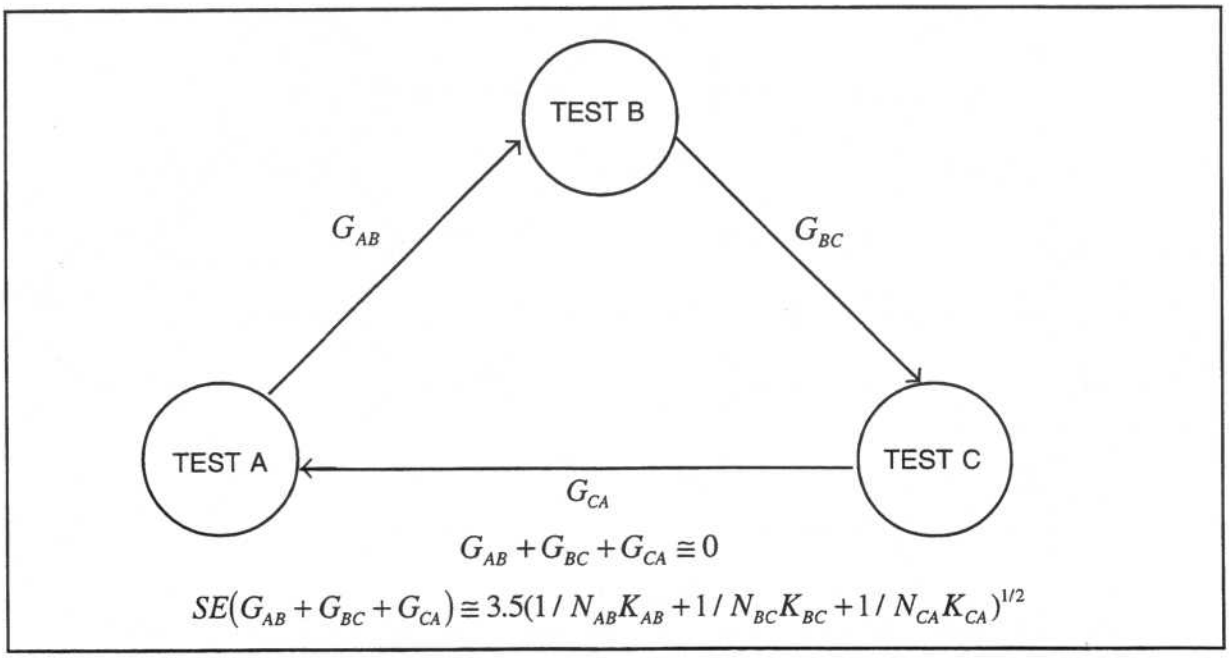


Each line connecting a circle represents a link of common items shared by the two tests it joins. Tests increase in difficulty horizontally along the variable and are comparable in difficulty vertically.

Links can be constructed to form a *loop* as shown in Figure 11.5.

Figure 11.5

A loop of three links.



The loop is an important linking structure because it yields an additional verification of link coherence. If the three links in a loop are consistent, then the sum of their three link translations should estimate zero.

$$(G_{AB} + G_{BC} + G_{CA}) \cong 0$$

where G_{AB} means the shift from Test A to Test B as we go around the loop so that G_{CA} means the shift from Test C back to Test A.

Estimating zero statistically means that the sum of these shifts should come to within a few standard errors of zero. The standard error of the sum $(G_{AB}+G_{BC}+G_{CA})$ is:

$$3.5(1/N_{AB}K_{AB} + 1/N_{BC}K_{BC} + 1/N_{CA}K_{CA})^{1/2}$$

in which

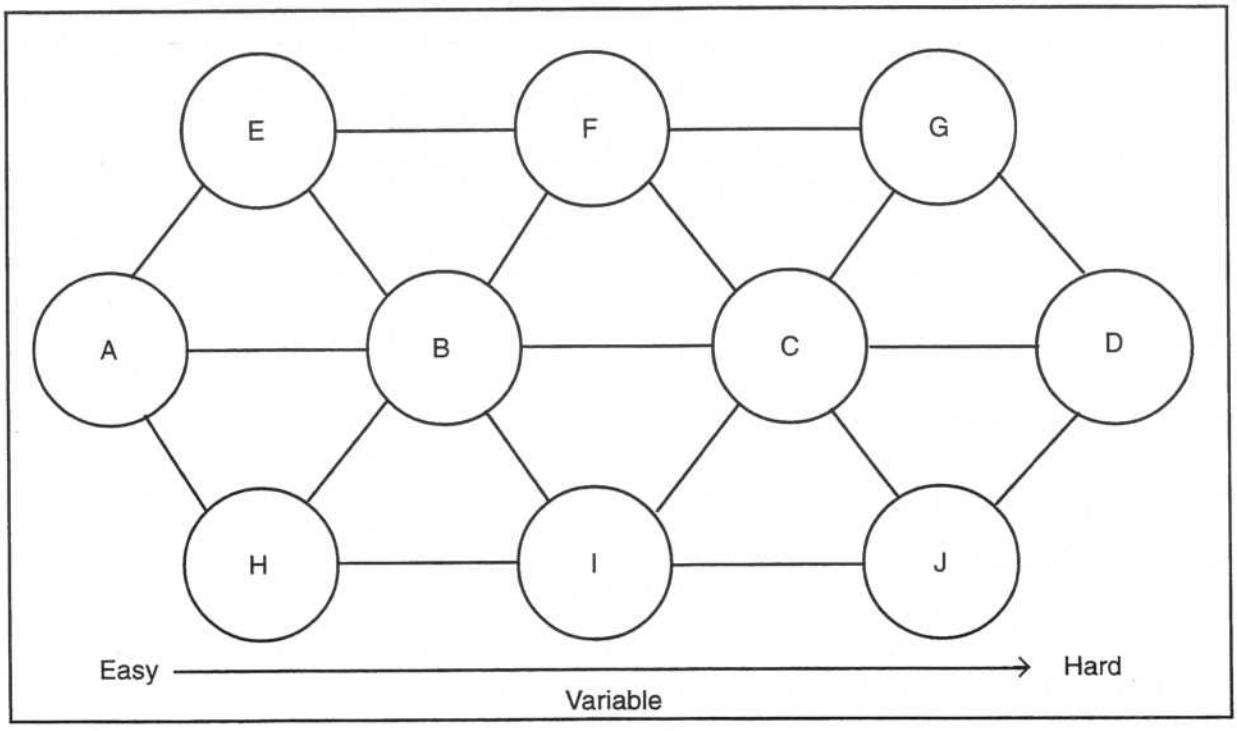
N = the various calibration sample sizes and

K = the various numbers of items in each link.

With four or more tests we can construct *networks* of loops. Figure 11.6 shows ten tests marking out several levels of difficulty from Tests A through D. This network could connect ten 60-item tests by means of nineteen 10-item links to construct a bank of 600-190=410 commonly calibrated items. If 100 persons took each test, then 410 items could be evaluated for possible calibration together from the responses of only 1,000 persons. Even persons at 50 per test would provide a substantial purchase on the possibilities for building an item bank out of the best of the 410 items.

Figure 11.6

A network connecting ten tests with nineteen links.



The building blocks of a test network are the loops of three tests each. When a loop fits the Rasch model, then its three translations will sum to within a few standard errors of zero. The success of the network at linking item calibrations can be evaluated from the magnitudes and directions of these loop sums. Shaky regions can be identified and steps taken to avoid or improve them.

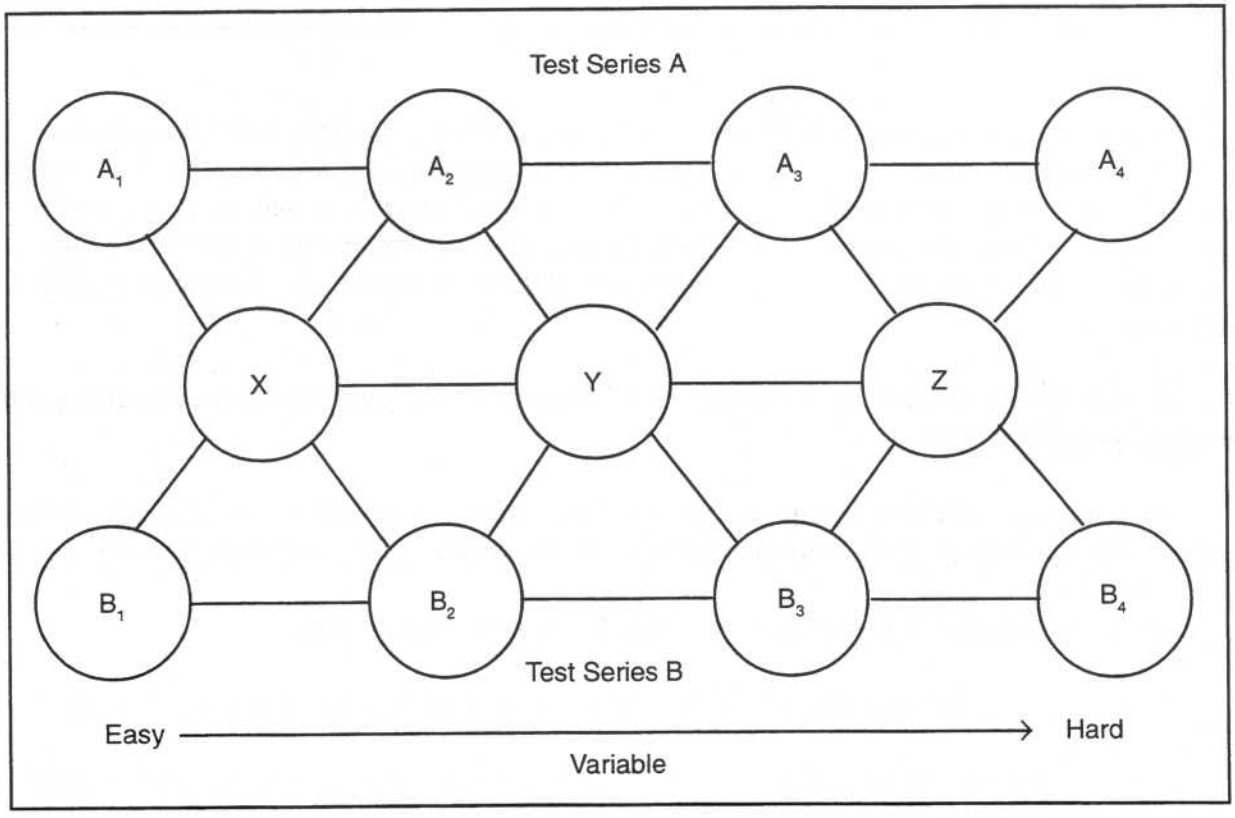
The implementation of test networks leads to banks of commonly calibrated items far larger in number and far more dispersed in difficulty than any single person could ever handle. The resulting item banks, because of the calibration of their items onto one common variable, provide the item resources for a prolific family of useful tests, long or short, easy or hard, widely spaced in item difficulty or narrowly focused, all equated in the measures they imply.

BANKING EXISTING TESTS AND ITEMS

These methods for building item banks can be applied to existing tests and items, if they have been carefully constructed. Suppose we have two non-overlapping, sequential series of tests A1, A2, A3, A4 and B1, B2, B3, B4 which we want to equate. All eight tests can be equated by connecting them with a new series of intermediate tests X, Y and Z made up entirely from items common to both series as shown in Figure 11.7.

Figure 11.7

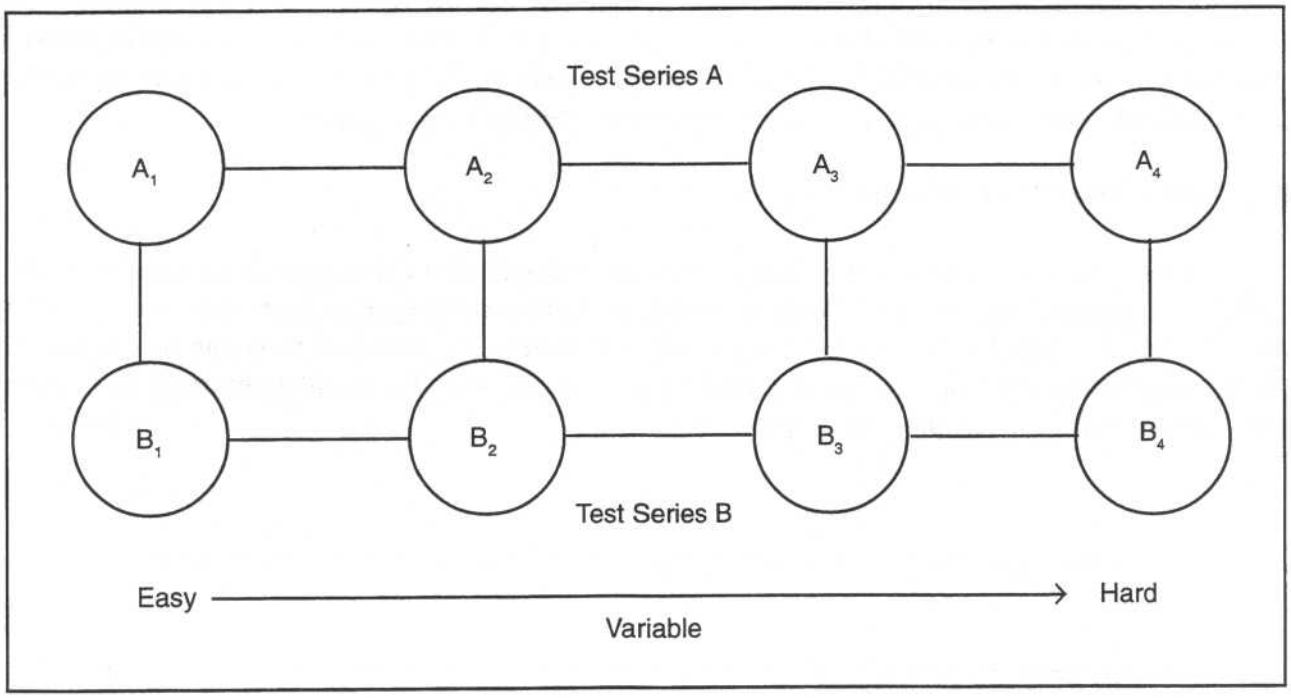
Connecting two non-overlapping test series by intermediate linking tests.



Were the A and B series of tests in Figure 11.7 still in the planning stage, they could also be linked directly by embedding common items in each test according to the pattern shown in Figure 11.8.

Figure 11.8

Connecting two test series by embedding common links.



Networks maximize the number of links among test forms because each form is linked to as many other forms as possible. To illustrate, take a small banking problem where we use 10 items per form in a *web* in which each one of these 10 items also appears in one of 10 other different forms. The complete set of 10+1=11 forms constitutes a web woven out of 11 x 10/2=55 individual linking items. Every one of the 11 forms is connected to every other form. The pattern is pictured in Figure 11.9.

The number entered in each cell is the identification of the item linking the two forms which define the position of that cell.

In this design, the web is complete because every form is connected to every other form. In the use of webs to build banks, however, there are three constraints which affect their construction:

1) the total number of items we want to calibrate into the bank,

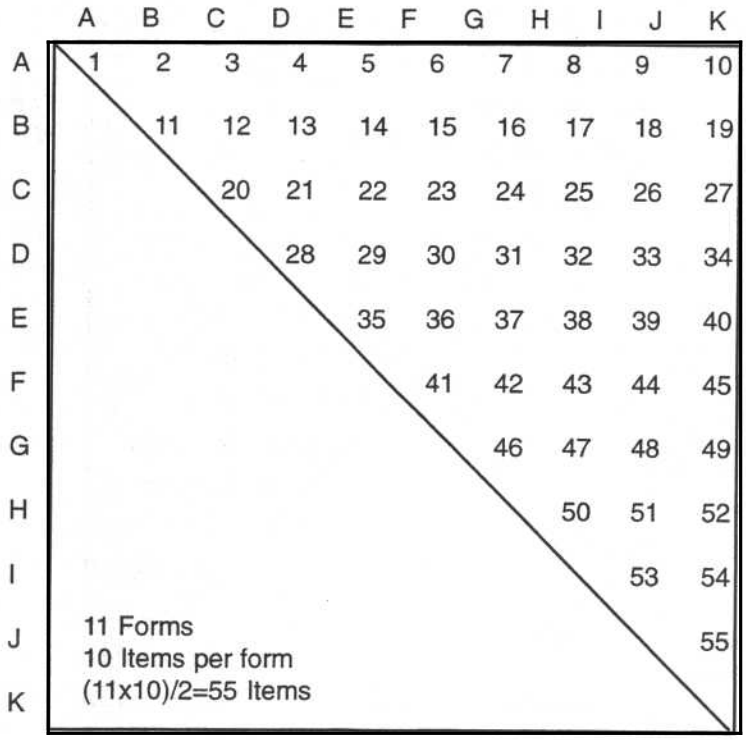
2) the maximum number of items which we can combine into a single form and

3) the extent to which the bank we have in mind reaches out in difficulty beyond the capacity of any one person.

The testing situation and the capacity of the persons taking the test forms limit the number of items we can put into a single form. Usually, however, we want to calibrate many more items that we can embed in a complete web like the one illustrated in Figure 11.9. There are two possibilities for including more items.

Figure 11.9

A complete web for parallel forms.

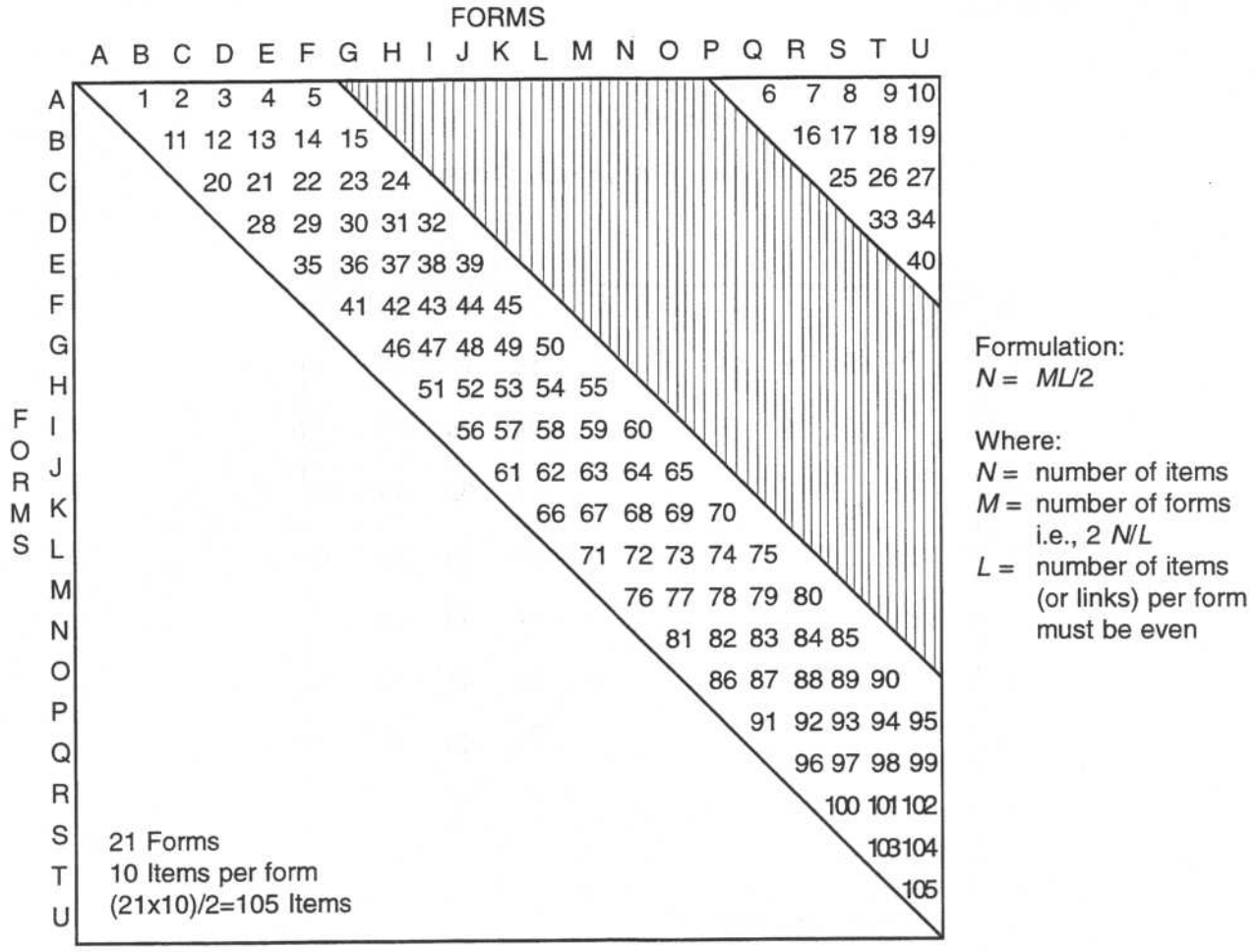


The simplest, but not the best, is to design a "nuclear" complete web which uses up some portion of the items we can include in a single form. Then we fill out the required form length with additional "tag" items. These tag items are calibrated into the bank by means of the link items in their form. Unlike the link items, however, which always appear in two forms, the tag items appear in only one form and so give no help with linking forms together into one commonly calibrated blank.

Another possibility, which is better statistically, is to increase the number of forms used while keeping the items per form fixed at the required limit. This makes the web incomplete but in a systematic way. The paired data on every item appearing twice can be used to evaluate the coherence of bank calibrations. Figure 11.10 shows an "incomplete" web for a 21 form design with 10 items per form, as in Figure 11.9, but connecting nearly twice as many items.

Figure 11.10

An incomplete web for parallel forms.

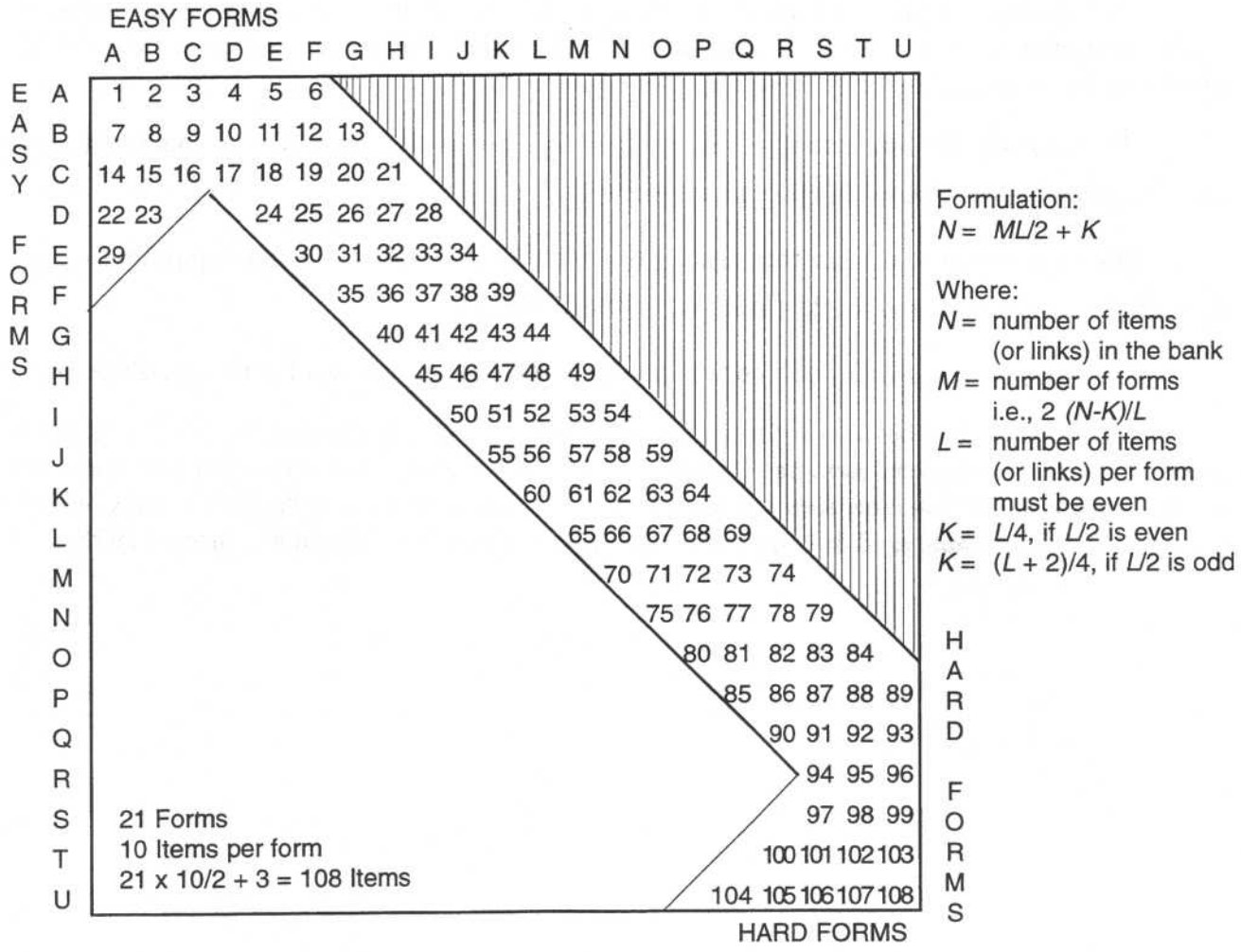


The incomplete web in Figure 11.10 is suitable for linking a set of parallel test forms. When the reach of the bank goes beyond the capacity of any one person, however, neither of the webs in Figures 11.9 and 11.10 will suffice, because we will be unable to combine items from the easy and hard ends of the bank into the same forms. The triangle of linking items in the upper right corners of Figures 9 and 10 will not be functional and will have to be deleted. In order to maintain the balance of linking along the variable we will have to do something at each end of the web to fill out the easiest and hardest forms so that the extremes are as tightly linked as the center.

Figure 11.11 shows how this can be done systematically for a set of 21 sequential forms. We still have 10 items per form, but now only adjacent forms are linked together. There are no common items connecting the easiest forms directly with the hardest forms. But over the range of the variable the forms near to one another in difficulty level are woven together with the maximum number of links.

Figure 11.11

An incomplete web for sequential forms.



Each linking item in the webs shown in Figures 11.8, 11.9, 11.10, and 11.11 could in fact refer to a cluster of two or more items which appear together in each of the two forms they link. Sometimes the design or printing format of items forces them into clusters. This happens in reading comprehension tests where clusters of items are attached to reading passages. It also occurs on math and information retrieval tests where clusters of items refer to common exhibits. Clustering increases the item length of each form by a factor equal to the cluster size.

The statistical analysis of a bank-building web is simple if the web is complete as in Figure 11.9. The row means of the corresponding matrix of form links are least square estimates of the form difficulties. We need only be careful about signs. If the web cell entry $G_{jk.}$ estimates the difference in difficulty $(\delta_i - \delta_k)$ between forms j and k and the form difficulties are centered at zero so that $\delta\cdot = 0$, then

$$G_{j\cdot} = \sum_{K}^{M} G_{jk} \,/\, M \approx \delta j$$

the row means of the link matrix calibrate the forms onto their common variable. Once form difficulties are obtained, they need only be added to the item difficulties within forms to bring all items onto the common variable shared by the forms.

The incomplete webs in Figures 11.10 and 11.11 require us to estimate row means from a matrix with missing data. The skew symmetry of link matrices helps the solution to this problem which can be done satisfactorily by iteration or regression.

When cells of the link matrix of $G_{jk.}$ are missing, then initial values for $G_{j.}$ can be obtained from Equation [6] by using zero's for the missing $G_{jk.}$.

The next step is to replace the missing $G_{jk.}$ with estimates from the corresponding $G_{j.}$ and $G_{k.}$ using $G_{jk} = G_{j.} - G_{k.}$ and recalculating $G_{j.}$ by Equation [6].

Iterations of this process will converge to stable values for the test form difficulties $G_{j.}$.

An even simpler but less informative solution is to express the data for all forms in one large matrix in which every item has its own column, every person has their own row and every intersection, at which a person does not address an item, is recorded as blank. This matrix, with its missing data, can be analyzed directly in one step with BIGSTEPS (Wright & Linacre, 1996).

MEASUREMENT ESSENTIALS

2nd Edition

BENJAMIN WRIGHT

MARK STONE

WIDE RANGE, INC.
Wilmington, Delaware