13. ITEM BANKING

This chapter discusses the curricular implications of item banking and its usefulness to all who depend on tests to evaluate educational achievement. We review the psychometric basis of item banking and give equations for building a bank. We conclude by showing how item quality control can be maintained over a bank of items.

THE IDEA OF ITEM BANKING

A mere collection of items is not an item bank. An item bank is a set of carefully composed and jointly calibrated items that develop, define and quantify a single common theme and hence provide an operational definition of one variable.

The first step in building an item bank is to develop its specifications. If we are building a scholastic variable it will be necessary to define the curriculum area and then to determine which items explicate it. To do so requires the expertise of professionals familiar with that curriculum area: teachers and curriculum experts.

We need a plan for the scholastic variable which is sufficiently detailed to specify how the items are expected to be ordered by difficulty along one main line of scholastic growth. This is important because it is in this beginning step that we demonstrate our understanding of the line of inquiry that is intended to define the scholastic variable under construction. If we discover that we do not have a clear enough understanding of the items to arrange them by difficulty order, then we have discovered that we do not know enough about what we are trying to do to succeed.

To accomplish item development:

1. Choose or write an item that you consider clearly on the line of the scholastic variable to be constructed.

   Intended Difficulty: ----1----> Harder

2. Add a second item written to be easier than the first item.

   Intended Difficulty: ----2----1----> Harder

3. Add a third item written to be harder than the first item.

   Intended Difficulty: ----2----1----3----> Harder

4. Next, add a fourth item positioned between items one and two and a fifth positioned between items one and three.

   Intended Difficulty: ----2----4----1----5----3---> Harder

5. Continue this stepwise process by positioning successively easier and harder items which extend the line of existing items and by filling in the spaces between these items with additional items positioned in difficulty between pairs of existing items.

This process of constructing the variable with items can be refined by re-positioning items upon further consideration and by review by other experts. The final line of items should show an ordering of items positioned by their intended difficulty from the easiest to the hardest. Successful construction of such a line of ordered items is an indication that the essence of the variable is understood by the item writers, and that the growth line implied by the scholastic variable and the items which define it belong together and lead somewhere. When we are not able successfully to position items along a line of growth by their difficulty, that is a sign that we do not understand our idea of the variable or the items required to describe it well enough to proceed.

Each item must represent an element in the strand of the scholastic variable we are building and each item must test some knowledge, skill or behavior at a specified position along the increase of that variable. When the items are empirically calibrated, these "conceptual" positions can be verified and improved. When, finally, the items are well-located along the line of a scholastic variable, then the scholastic variable has acquired a meaningful and useful operational definition.

Items with low calibration values entail easy tasks that define the low end of the variable. Items with high calibration values entail difficult tasks that define the high end of the variable. The arrangement of items by their order of calibrations from easy to hard describes the path of learning that most students follow as they progress along the line of the scholastic variable. The empirical item calibrations can be obtained by applying the Rasch model for what ought to happen when a student attempts an item (Rasch, 1960/1980; Wright & Stone, 1979). This probability model imposes an orderly response process on the data. The probabilities obtained specify what is expected to occur, with some give and take, because no student will follow the expected line exactly.

The process of item planning, writing and positioning, along with the confrontations and revisions provoked by subsequent item calibrations, is an integrated and constructive dialogue between the item construction phase of bank development and the item calibration phase - between theory and practice. This dialogue will progress in successive stages as better and better confirmation of item positions is achieved and the operational definition of the scholastic variable evolves. Continual monitoring of the bank building process is both required and beneficial.

When a scholastic variable is well understood, the task of constructing its item bank is straightforward. But when the variable is newly conceived or not clearly understood, the interactive process between item positioning and item calibration may require many stages before useful agreement between intention and realization, between idea and experience is achieved.

It is important to recognize that the agreement to be achieved between theory and practice is the method for control over item development quality. Creative item writing is required to capture and implement the essence of a scholastic variable. The empirical calibration of these items gives the item writers feedback on the utility of their creative efforts.

Reviewing the evolving line of items from easy to hard along the intended variable promotes communication between the specialists of curriculum and teaching and those of test construction. The resulting marriage of these two specialty areas can produce valid scholastic variables defined by operationally efficient items.

THE USEFULNESS OF ITEM BANKS

A well constructed and organized item bank enables a wide variety of tests. Each test can be tailored to the objectives of its use and yet be quantitatively connected to the common core of bank items. Additional items can be added whenever their calibrations are found to fit the growing common core of calibrated items.

A well constructed item bank provides the elements necessary for designing the best possible test for any assessment purpose. It is not necessary for every student to take the same test in order to be able to compare results. Students can take only those items closest to their level of development as in computer assisted instruction. The number of items, level, range of difficulty and content can be selected individually from the bank. Each individualized test maintains quantitative comparability because any test formed from calibrated bank items, on which a valid pattern of performance is obtained, can be automatically equated through the calibration of the test items to all of the items in the bank and so to all of the measures produced by every other test that has ever been or might sometime be formed from this bank.

A very wide-range test for general screening can be formed as well as narrow tests for specific purposes. The two procedures of wide-range screening and narrow-range measuring can be combined to implement adaptive testing. The wide-range test locates the student's general area on the line of the scholastic variable and the narrow-range test pinpoints the location for the most efficient measurement of that student.

BUILDING AN ITEM BANK

TO CONSTRUCT AN ITEM BANK:

1. Begin with a pool of items dominated in their content by a common curriculum line. These items are best when constructed and arranged according to a clear hierarchy of increasing conceptual difficulty.

2. Apportion these items among test forms so that there is a web of common items which forms a network of connections among all test forms. This web can reduce the test size of each form to manageable length and yet distribute all items over the many forms connected by the web of shared items.

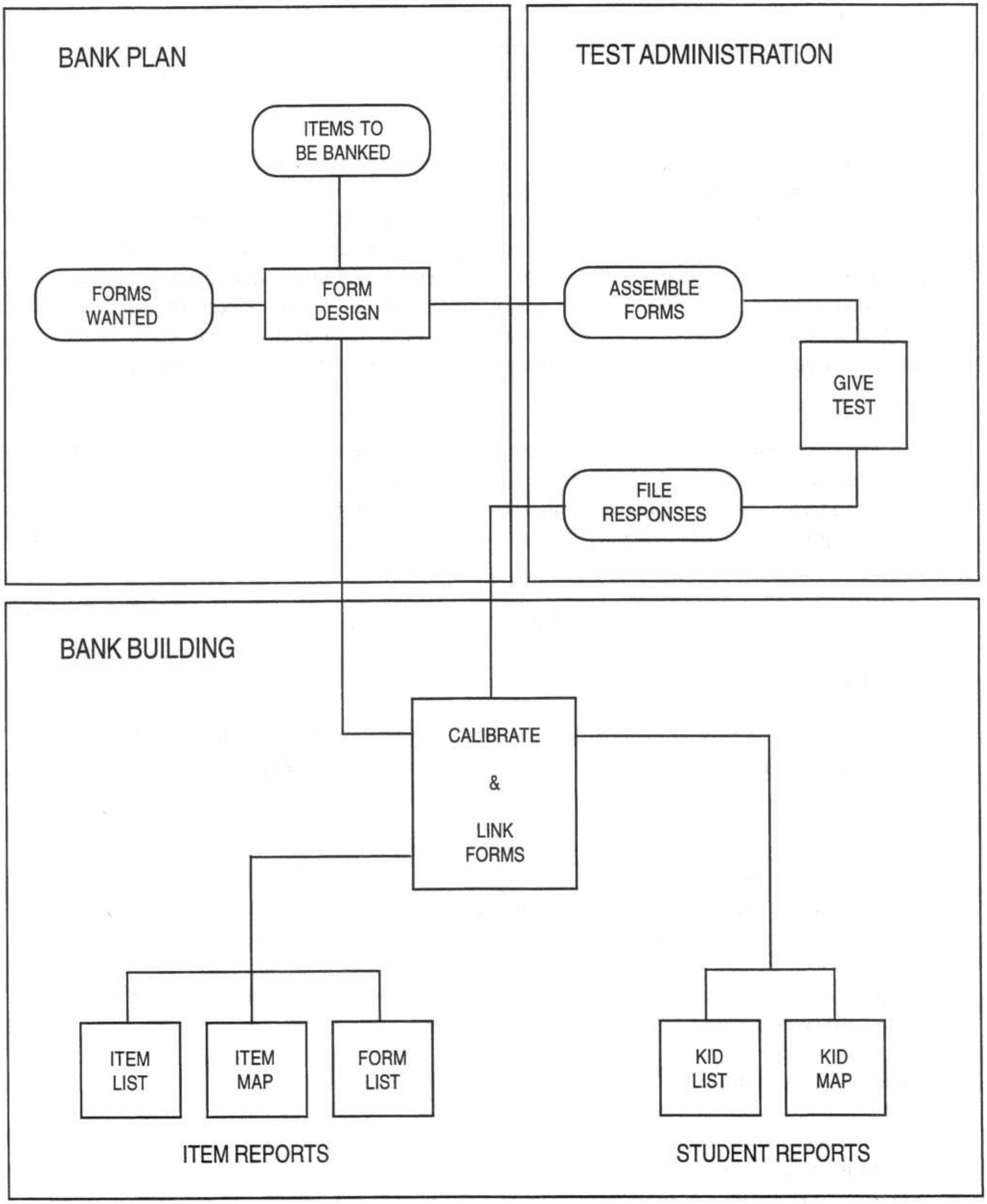
The flow chart in Figure 13.1 outlines the basic steps necessary to build a pool of coordinated items into a calibrated bank.

DESIGNING TEST FORMS

Items must be distributed among test forms so that there is a web of common item connections which maximizes the statistical strength of the linking structure, while meeting the practical requirements of the test situation (for details see Wright & Stone, 1979, Chapter 5).

Figure 13.1

Flowchart for building an item bank.



Design input includes the number of items to be calibrated, the number of items desired per form, the number of items desired per link, the expected difficulty of each item and whether the pattern of form difficulties is to be horizontal or vertical. The design determines the number of links per form, total number of links and total number of forms necessary for an optimal web.

The design process constructs a file of item specifications from which the banking system works. This list includes item identification number, name, link number, expected difficulty, correct responses, and associated forms so that item test form placements can be checked and listed item-by-form and also form-by-item in their within-form position in order to facilitate the verification of content coherence and form assembly.

CALIBRATING TEST FORMS

When forms are designed, assembled and administered, student responses are collected, recorded and filed in an individual record for each student that includes student identification, form taken, and the student's item response string. This student file is the form calibration input. The item file prepared during form design and the student file obtained from testing, are used to calibrate items within each form in order to analyze within-form item and student fit and then to calibrate all items and measure all students simultaneously on one common linear variable. (A useful computer program for this is *BIGSTEPS*, Wright & Linacre, 1997.)

The form equating, accomplished by the single simultaneous analysis of all forms, can be evaluated in detail by explicitly linking the separate analyses of each form in which item difficulties are still relative to the local origin defined by each form. Connections among forms can be made explicit by a link analysis of the connections of all forms to the single common scale.

ANALYSIS OF FIT

Analysis of fit evaluates the degree of consistency between observation and expectation and the extent to which any subdivisions of observed data (by group, grade level, sex, etc.) produce statistically equivalent item and form calibrations. There is a hierarchy of fit statistics available to implement fit analysis.

ITEM WITHIN-FORM FIT

A routine check on whether item difficulties are sample-free is done during form calibration. If item estimates are invariant with respect to student abilities, student sample subdivisions will give statistically equivalent item difficulties. One way to evaluate sample-freeness is to divide the sample into raw score subgroups and then to compare the observed successes on each item *i* in each raw score subgroup *g* with the number of successes predicted for that subgroup. If the general parameter estimates are adequate for describing score group *g*, then the observed number correct in group *g* will be near the estimated model expectation

$$R_{gi} = \sum_{r \in g} N_r p_{ri} \qquad 13.1$$

with model variance

$$s_{gi}^2 = \sum_{rEg} N_r p_{ri}[1 - p_{ri}]$$

where 13.2

$$p_{ri} = \frac{\exp(b_r - d_i)}{1 + \exp(b_r - d_i)}.$$

N_r is the number of students with raw score *r* and p_{ri} is the estimated probability of success for a student with score *r* on item *i*, given the general ability and difficulty estimates b_r for score *r* and d_i for test item *i*.

If observed and expected numbers correct are statistically equivalent, given the model variance of the observed, then there is no evidence against the conclusion that the subgroup concurs on the estimated difficulty of item *i*. The statistical precision (reliability) of this estimate can be specified with its modeled standard error. Similar analyses can be done for student subgroups defined in other ways.

Another way to check within-form item fit is to evaluate the agreement between the variable manifested by item *i* and the variable defined by the other items. A useful statistic for this is an "infit" mean square in which the standard squared residual of observation *x* from its expectation *p*, $z_{ni}^2 = (x_{ni} - p_{ni})^2 / [p_{ni}(1 - p_{ni})$, for each student *n*'s response to item *i*, is weighted by the information in the observation, $q_{ni} = p_{ni}(1 - p_{ni})$, and summed over the *N* students.

$$V_i = \frac{\sum_n^N [z_{ni}^2 q_{ni}]}{\sum_n^N q_{ni}} [N / (N - 1)] \quad . \qquad 13.3$$

This "infit" statistic is useful because it is robust with respect to idiosyncratic outliers. The alternative "outfit" statistic that detects outliers is the unweighted mean square,

$$U_i = \sum_n^N z_{ni}^2 / (N - 1) \ . \qquad 13.4$$

When data fit the model, these statistics estimate one with variance of order $[2 / (N - 1)]$.

For more exact estimates of these variances see Rasch, 1980, pp. 193-194 or Wright & Masters, 1982, p. 100.

CALCULATING TEST FORM LINKS

When the items in each form have been calibrated separately within each form, there are as many difficulty estimates for each item as there are forms in which it appears. The items that appear in more than one form provide the linking data. The differences observed between within-form item calibrations and the model requirement that each item be characterized by a single difficulty, regardless of form or sample, estimate the relative difficulty of each form. This form difficulty is then added to the within-form item calibrations to place every administration of every item onto one common bank scale.

CALIBRATING FORMS ON THE BANK

To estimate the shift in difficulty between two forms, k and j, a weighted average of difficulty differences is calculated for the items linking them

$$t_{kj} = \frac{\sum_{i}^{n}[d_{ij} - d_{ik}]w_{ikj}}{\sum_{i}^{n} w_{ikj}} \qquad 13.5$$

where d_{ik} and d_{ij} are the estimated difficulties of linking item i in forms k and j, n is the number of items in this link, and $w_{ikj} = 1/(se_{ik}^2 + se_{ij}^2)$ is an information weight based on the item calibration standard errors, se_{ik} , and se_{ij} . The standard error of the difficulty shift t_{kj} is

$$se_{kj} = 1/\left[\sum_{i}^{n} w_{ikj}\right]^{1/2} . \qquad 13.6$$

The shift t_{kj} estimates the difference in origins of forms k and j. A shift is calculated for every pair of forms linked by common items. When every possible pair of forms is linked, then the difficulty T_k of form k is the average shift for form k over all forms.

$$T_k = \frac{\sum_{j}^{M} t_{kj}}{M} \qquad 13.7$$

where M is the number of forms and $t_{kk} = 0$. The standard error of form difficulty T_k is

$$se_k = \left(\sum_{j}^{M} {se_{kj}}^2\right)^{1/2} / M \qquad 13.8$$

Equations 13.5 through *13.8* assume every form is linked to every other form. When links are missing between some forms, as is usually the case, an iterative procedure can be used to bridge the empty cells. Empty cells can be started at

$$t_{kj} = t_{jk} = 0 \qquad 13.9$$

and the form calibrations T_k improved step-by-step by calculating temporary form difficulties with *Equation 13.7*, adjusting empty cells to

$$t_{kj} = T_k - T_j \text{ and } t_{jk} = T_j - T_k \qquad 13.10$$

and then reapplying *Equation 13.7* iteratively until the successive values of T_k stabilize. This process works as long as every form can be reached from every other form by some chain of links.

This procedure sets the bank origin at the center of all forms so that form difficulty T_k is the difference between the center of form k and the center of the bank.

ITEM WITHIN-LINK FIT ANALYSIS

To verify the extent to which the linking items perform adequately within their forms combine the item-within-form fit statistics of *Equation 13.3* into a within-form fit statistic for the link.

$$\text{Within form link fit} = \frac{\sum_{i}^{n}(V_{ik} + V_{ji})}{2n} \qquad 13.11$$

where V_{ik} is the fit of item i in form k

V_{ij} is the fit of item i in form j, and

n is the number of items in the link.

This statistic estimates one with variance of order $[1/n(N-1)]$ when the link items fit within forms.

ITEM BETWEEN-LINK FIT ANALYSIS

To check the extent to which link items agree on the relative difficulties of their two forms, calculate the ratio of observed to model variance.

$$\text{Between form link fit} = \frac{\sum_{i}^{n}(d'_{ik} - d'_{ij})^2}{\sum_{i}^{n} w_{ikj}} \qquad 13.12$$

where now $w_{ikj} = [se_{ik}^2 + se_{ij}^2]$ and the within form item difficulties, d_{ik} have been translated to their bank values d'_{ik} by

$$d'_{ik} = d_{ik} + T_k \qquad 13.13$$

Values substantially greater than one, given expected variance $[2/(n-1)]$, signify that some items operate differently in the two forms. A plot of d_{ik} versus d_{ij} over i facilitates the evaluation of link status and the identification of aberrant items (see Wright & Stone, 1979, pp. 92-95; Wright & Masters, 1982, pp. 114-117).

LINK WITHIN-BANK FIT ANALYSIS

To check the extent of agreement among links with respect to form difficulties review the extent to which each entry in the matrix of observed shifts between forms is close to the difference in bank difficulties of the forms. To evaluate whether a link fits the bank, calculate the link residual

$$y_{kj} = t_{kj} - [T_k - T_j] \qquad 13.14$$

where t_{kj} is the observed shift between forms *k* and *j*, and T_k and T_j are their bank difficulties.

These link residuals can be standardized to mean zero and variance one by dividing them by the standard errors, se_{kj} of their t_{kj} of *Equation 13.5* and multiplying by $[M/(M-2)]^{1/2}$ where *M* is the number of forms in the linking network.

FORM WITHIN-BANK FIT ANALYSIS

To check the fit of each form to the bank as a whole calculate

$$V_k = \frac{\sum_{i}^{L}[y_{kj} / se_{kj}]^2[M/(M-2)]}{L-1} \qquad 13.15$$

where *L* is the number of t_{kj} observed for form *k*.

The criterion value of V_k is also one, this time with variance of order $[2/(L-1)]$.

The fit of a link or a form into the bank is related to how well linking items fit within their own forms. When the number of students taking a form is large, the item fit statistic variances can become unrealistically small and must be taken with a grain of salt. Careful investigation of doubtful items is always instructive and invariably leads to insight into the nature of the variable. The misfit of links within the bank is usually associated with particular forms. This can occur when a form is inadvertently administered to a sample of students for whom it is inappropriate. The best items for estimating form difficulties are those that satisfy the various fit analyses.

REVIEWING THE RESULTING BANK

At this point an ITEM LIST (Figure 13.1) which gives each item in the bank by sequence number, legitimate alternatives, correct responses, item name, bank difficulty, between difficulty root mean square, and within form fit mean square is useful.

Bank difficulty is the average of the item's difficulties in the forms in which it was calibrated, adjusted for these forms' local difficulties. A between difficulty root mean square, the square root of the average squared difference between an item's bank equated difficulties in each form and its bank difficulty is useful to tag potentially errant items. Items showing between difficulty root mean squares greater than 0.5 logits are frequently found to have been miskeyed or misprinted in one of the forms in which they appear.

The within-form item fit mean square of *Equation 13.3* can be standardized to mean zero and variance one so that the average square of these standardized within-form fits can summarize item performances within forms. Its sign is taken from the sign of the standardized fit with the largest absolute value to distinguish between misfit caused by unexpected disorder, indicated by large positive standardized fits, and misfit caused by unexpected within-form inter-item dependence, indicated by large negative standardized fits. It is useful to tag items producing values greater than 2 or less than -2 for further examination.

An ITEM MAP (Figure 13.1) which displays the variable graphically by plotting the items according to their bank difficulties along the line of the variable which they define, will enable teachers to examine the relationship between the content of the items and their bank difficulties in order to review the extent to which the empirical item order defines a curriculum strand that agrees with their curriculum expectations and so has construct validity for them. The item map provides a framework for writing new items to fill gaps that appear in the definition of the curriculum strand and for choosing items for new tests.

A FORM LIST (Figure 13.1) which gives each form by form number, name, number of items and bank difficulty is useful. Each item is listed by form position, item name, key, within form difficulty and standard error, total within form standardized fit, and bank difficulty. This facilitates the review of each form as a whole and the identification of form specific anomalies.

A KID LIST (Figure 13.1) which gives each student by identification, ability measure, error and fit statistic indicates which students misfit by displaying their response string and its residuals from expectation, so that teacher and student see the specific item sources of misfit.

A KID MAP (Figure 13.1) produces a graphical representation of each individual student's performance. The map for each student shows where that student and the items they took stand on the curriculum strand, which items were answered correctly, the probability of each response, and the student's percent mastery at each item. This provides teacher, student and parent with a picture of the student's performance which combines in one easy to read picture specification of criteria mastery with the identification of unexpected strengths and weaknesses.

ITEM QUALITY CONTROL

Once items have been banked, the identification and study of misfitting items follows. The irregularities most often identified are mechanical and clerical such as miskeying, misprinting, misscoring, more than one right answer and no right answer. Sometimes, however, item misfit brings out anomalies in student performance which leads to new and unexpected understanding of how the subject matter contained in the item is learned and used.

The item infit and outfit mean square fit statistics of *Equations 13.3* and *13.4* indicate the degree to which an item functions as intended. Mean square statistics greater than 1.4 imply noise in item use, outbreaks of guessing or carelessness, or the presence of secondary variables correlated negatively with the intended variable. Mean square statistics less than 0.6 imply inter-item dependencies or the presence of secondary variables correlated positively with the intended variable.

MISFIT PATTERNS

Miskeying and scanner errors usually cause an item to appear more difficult than anticipated, making item fit too large.

Misfit caused by student behavior, such as guessing and carelessness, is not diagnosed well by item fit statistics because item statistics lump together students behaving differently. Disturbances that are the consequences of individual student behavior are best detected and best dealt with through the fit analysis of individual students (Wright & Stone, 1979, Chapters 4 & 7). But item statistics can call attention to items that tend to provoke irregular behavior in many students.

Guessing is a problem only when students inclined to guess are also provoked to guess on items that are too difficult for them and then only when those particular students happen to guess correctly. This is more probable for low ability students but may occur for others depending upon the value given to an outcome or success on the test and the time allowed. Problems of guessing are best addressed by targeting test administration so that it does not provoke guessing by allowing enough test time so that students are not rushed and by reviewing each student's response pattern for the presence of improbable right answers which might have been achieved by lucky guessing.

Carelessness occurs when a high ability student fails an easy item. The pattern in item statistics is low difficulty and high fit. This, too, is most usefully and accurately detected through the identification of improbable wrong answers in individualized person fit analyses.

OTHER SOURCES

When the disturbance in a misfitting item is not mechanical or clerical, the cause is usually special knowledge. Interactions with curriculum specifics affect the shape of the response curve. Dependence on a skill that only high-ability students are taught can make an item unfairly easier for these high ability students. This will cause the item to have a fit statistic that is improbably low and an unusually high point biserial. On the other hand, dependence on a skill that is negatively related to instruction, so that low-ability students possess more of it, can make an item unfairly easier for low-ability students and, hence, give it a fit statistic that is improbably high. Either way, the interaction disqualifies the item for use with students who are unequal in their exposure to the special skill. When fit is too high, the item is unfair to more able students. When fit is too low, the item is unfair to less able students.

One-step implementation of an item bank can be done using a computer program like *BIGSTEPS* (Wright & Linacre, 1997). But the data layout must be organized so that the separate forms flow into one standard frame of reference. An integrated item banking system like *SAMS* (Wright, Linacre & Schultz, 1991) can be used for general school applications.

MEASUREMENT ESSENTIALS

2nd Edition

BENJAMIN WRIGHT

MARK STONE

WIDE RANGE, INC.
Wilmington, Delaware