17. INFORMATION AND MISFIT ANALYSIS

THE STANDARD ERROR OF A MEASURE

The most immediate and also most convenient quantification of the precision of an empirical measure is the standard error (SE) of the measure's estimate. The magnitude of the SE is either determined from the factual structure of the measuring instrument (as in "to the nearest sixteenth of an inch" on a ruler) or calculated from the measurement model used to calibrate the instrument. It is usually estimated from the same data used to estimate the measure. This SE estimates the standard deviation of innumerable independent replications of the data collecting process, when the only disturbances imagined are those anticipated by the measurement model.

The convenience of the SE quantification of precision is that it is in the units of the measure and so can be used directly to specify:

1) an "identification of misfit", as in outside three standard errors (SE);

2) a "region of confidence", as in within two standard errors (SE);

3) an "allowance for error", as in one standard error (SE).

The inconvenience of the SE is that when several samples of independent data bearing on a common quantity to be estimated are combined to form a "better" estimate or when it is useful to keep track of the sequential improvement of "precision" during a stepwise process of data collecting, the corresponding SE's are not additive.

INFORMATION

Ronald Fisher devised a cure for this inconvenience in the 1920's (e.g. 1935, p. 182 ff.). While the SE's of a series of independently obtained, but commonly bearing, commensurable measures are not additive, their inverse squares are. When applied to Rasch measurement "Fisher information" can be defined as

$$I = C / SE^2$$

where C = a constant chosen to specify convenient "information" units.

For dichotomous data, as in test item responses scored 0 or 1, the inverse square of each measure's SE is proportional to a count of how many "standard" items inform that measure.

In particular, when $C \equiv 4$, then

$$I = 4 / SE^2$$

becomes the minimum number of perfectly targeted (i.e. maximally informative) items it would take to produce this SE. I is the "information" in the estimate. We will call the units of $I = 4 / SE^2$ "EQUITS," for EQUivalent on-target ITemS.

The additivity of $I = 4 / SE^2$ can be seen in the algebraic definition of SE for Rasch modeled dichotomous data.

Then

$$SE^2 = 1 / \left[\sum_i^L P_i (1 - P_i) \right]$$

where

$$P_i = \exp(B - D_i) / \left[1 + \exp(B - D_i) \right]$$

is the probability of a right answer given person measure B and item difficulty D_i, and $\sum_i^L$ signifies summation over the L items taken. Thus

$$I = \left(4 / SE^2 \right) = 4 * \sum_i^L \left[P_i (1 - P_i) \right]$$

is an expression which adds $\left[P_i (1 - P_1) \right]$ over items.

When every item is perfectly targeted, then

$$P_i = 1 / 2,$$

$$P_i (1 - P_i) = 1 / 4$$

and so

$$I = 4 * \left[\sum_i^L P_i (1 - P_i) \right] = L$$

the number of responses to perfectly targeted items necessary in order to obtain this particular

$$SE = \left\{ 1 / \left[\sum_i^L P_i (1 - P_i) \right] \right\}^{1/2}$$

COMPARING INFORMATION

When we wish to compare the information value of a pair of measures, we can use their corresponding SE's and this definition of information, $I = (4 / SE^2)$, to find out which measure contains more information and by how many "equivalent on-target items" or "equits".

Thus for measures B_1 and B_2 we have

$$I_1 = 4 / SE_1^2 \text{ equits}$$

$$I_2 = 4 / SE_2^2 \text{ equits}$$

for which the advantage in equits of measure B_2 over measure B_1 is

$$I_2 - I_1 = 4\left(1 / SE_2^2 - 1 / SE_1^2\right)$$

$$= 4(SE_1^2 - SE_2^2) / SE_1^2 * SE_2^2$$

Comparison of the information values of a pair of measures can also be calculated from the ratio of their error variances,

$$RE_{21} = I_2 / I_1 = SE_1^2 / SE_2^2$$

This ratio, RE_{21}, gives the "information" provided by the second measure B_2 in units of the "information" provided by the first B_1, i.e., it is the "Relative Efficiency" of the second measure with respect to the first.

MISFIT ANALYSIS

Maximum information is obtained when $P = .5$ so that $P(1 - P) = .25$. While this would appear ideal, there is a catch. Fit analysis requires the possibility of improbable, and hence unexpected responses - responses for which

$$P \rightarrow 1 \text{ but } X = 0 \text{ or } P \rightarrow 0 \text{ but } X = 1.$$

Then, when the highly probable response is *not* observed, misfit and hence invalidity is implied. Were all items targeted successfully near $P = .5$, this kind of fit analysis for verification of response validity would not be possible. Since $X = 0$ or 1 would be equally likely, no improbable condition with which to detect misfit could be observed.

THE EFFICIENCY - FIT PARADOX

1. Responses to items which provide maximum information because $P \rightarrow .5$ allow minimum misfit detection.
2. Responses to items which allow maximum misfit detection because $P \rightarrow 0$ or 1 provide minimum information.

Best test design requires a compromise between these extremes. The simultaneous avoidance of both extremes benefits greatly from prior knowledge concerning the relative locations of items and persons.

BEST TEST DESIGN

We want the items to elicit maximum information from the person. But we must balance the amount of information (reliability, precision) gained against the concomitant loss of opportunity to detect misfit and, hence, to verify validity. Where we have no knowledge of a person's ability, then items must be of a difficulty range sufficiently wide to cover the reasonable possibilities. This means that while some items will identify the location of the person between items passed and items failed, other items will inevitably turn out to be too far from the person's discovered ability to contribute much information about that ability. The off-target items, however, will be useful for identifying misfit and thus verifying validity.

When we have a useful expectation about where a person is on the variable to be measured, then item selection can be accomplished with maximum utility and efficiency by focusing most of the items on the interval in which we expect the person to be located, but including some additional intentionally off-target items to verify the validity of this location.

We use enough targeted items to "fix" the person's location with sufficient precision ($SE^2 \approx 4 / L$), where this L is the number of on-target items) for our testing purpose. Then we add enough additional off-target items (2 logits above and below where we expect the person to be located) to verify the validity of our measure.

The efficiency of this design depends on the extent of our knowledge of the person prior to the test. Without some prior focusing knowledge, we must use a wide range of items. This will guarantee enough off-target items to validate the measure, but will cost more items than a narrow on-target test to reach equivalent precision.

Targeting an educational test to a particular student requires both the art of knowing the student and the science of measurement. Teaching intuition can guide expectations in the absence of quantitative knowledge. When previous measurements are also available, they too can be utilized.

INFORMATION, EFFICIENCY AND PRECISION

The way information and efficiency enter into judging the value of an observation is through their bearing on the precision of measurement. Measurement precision depends on the number of items in the performance and on the difference in logits $|B - D|$ between each item difficulty and the person's ability. We can simplify the evaluation of each item's contribution to our knowledge of the person by calculating what percent of a best possible item the item in question contributes. These are the values of INF entered in Column 2 of Table 17.1.

We call this information index $INF = 400[P(1-P)]$

the "relative efficiency" of the observation.

The relative efficiency (INF) is the I defined in *Equation 17.8* but scaled by the factor 100 so

Table 17.1

Information and Misfit Statistics

1	2	3	4
LOGIT DISTANCE BETWEEN PERSON AND ITEM	% EFFICIENCY OF AN OBSERVATION AT $\lvert B-D \rvert$	NUMBER OF ITEMS L NEEDED TO MAINTAIN EQUAL PRECISION	IMPROBABILITY OF AN UNEXPECTED ANSWER AT $\lvert B-D \rvert$
$\lvert B-D \rvert$	$INF = 400P(1-P)$	$L = 1000 / INF$	$P = 1/\left[1+\exp(\lvert B-D \rvert)\right]$
0.0, 0.3	100	10	.50
0.4, 0.8	90	11	.33
0.9, 1.2	75	13	.25
1.3, 1.4	65	15	.20
1.4, 1.5	55	18	.17
1.7, 1.8	50	20	.14
1.9, 2.0	45	22	.12
2.1	40	25	.11
2.2	36	28	.10
2.3	33	30	.09
2.4	31	32	.08
2.5	28	36	.08
2.6	25	40	.07
2.7	23	43	.06
2.8	21	48	.06
2.9	20	50	.05
3.0	18	55	.05
3.1	16	61	.04
3.2	15	66	.04
3.3	14	73	.04
3.4	12	83	.03
3.5	11	91	.03
3.6	10	100	.03
3.7	9	106	.02
3.8	9	117	.02
3.9	8	129	.02
4.0	7	142	.02
4.1	6	156	.02
4.2	6	172	.02
4.3	5	189	.01
4.4	5	209	.01
4.5	4	230	.01
4.6	4	254	.01

Wright & Stone, 1979. *Best Test Design.* Chicago: MESA Press. Pages 73 and 216.

that it will give the amount of information provided by the observation at $|B-D|$ as a percentage of the maximum information that one observation "exactly on target" at $|B-D|=0$ would provide.

The relative efficiency (INF) of an observation can be used to estimate the potential value of any particular item for measuring a particular person. This can be done by considering how much information would be lost by removing that item from the test. Thus, $\text{INF}=23\%$ for $|B-D|=2.7$ indicates how much of a perfectly targeted item we gain by including that item in the measurement of the person and conversely how much we lose by omitting that item. The "how much" is 23% of the most we could get from one item exactly on target at $|B-D|=0$.

When an item and person are close to one another $|B\quad D|\quad 0$; i.e., on target, then the item contributes more to the measure of the person than when the item and person are far apart $|B\quad D|$. The greater the difference between item and person, the greater the number of items needed to obtain a measure of comparable precision and as a result, the less efficient each item.

Once we have estimates of person ability B to combine with our knowledge of item difficulty D, we can determine the relative efficiency of any item. Column 2 of Table 17.1 gives the percent relative efficiency (INF) by which any observation at the absolute difference $|B-D|$ given in Column 1, provides information about that person-item interaction.

It requires five $\text{INF}=20\%$ items at $|B-D|\rightarrow 2.9$ to provide as much information about a person as could be provided by one $\text{INF}=100\%$ item at $|B-D|\rightarrow 0$.

When $|B-D|$ is three, it takes four times as many items to equal the information to be had from items in the $|B-D|<1$ region, within one logit of the person.

The test length necessary to maintain a specific level of measurement precision is inversely proportional to the relative efficiency of the items used. The number L of less efficient items necessary to match the precision of 10 exactly-on-target, $|B-D|=0$, items is given in Column 3 of Table 17.1.

Column 3 shows $L=1000\,/\,INF$ the number of items needed to maintain equal precision over the range of possible values of $|B-D|$.

There is also, however, the verification or validation of test performance validity to keep in mind. When we are off-target because $|B-D|>2$ or 3, then we can use the possibility of unexpected (improbable) responses to evaluate response validity. Column 4 in Table 17.2 gives the probability of an unexpected response (i.e. the improbability of the observed response) for each value of $|B-D|$.

Note that as $|B-D|>2.8$, the probability of an unexpected response such as

$$X=0 \text{ when } (B-D)>2.8 \text{ or } X=1 \text{ when } (B-D)<-2.8$$

drops to P = .05. This produces the possibility of a statistically significant "misfit" and hence of a probable invalidity in that response to that item.

Detailed examples of misfit analysis are given in Chapter 4 of Best Test Design (Wright and Stone, 1979).

To standardize our use of Table 17.1, we use this guide:

Location of Item	(Ability-Difficulty) Difference	Item Efficiency and Misfit Detection
Right on Target	$\|B-D\|<1$	- excellent efficiency, 75% or better - no misfit analysis possible
Close Enough	$1<\|B-D\|<2$	- good efficiency, 45% or better - no misfit analysis possible
Slightly Off	$2<\|B-D\|<3$	- poor efficiency, less than 45% - misfit detectable when unexpected responses accumulate
Rather Off	$3<\|B-D\|<4$	- very poor efficiency, less than 18% - even single unexpected responses indicate irregularity
Extremely Off	$4<\|B-D\|$ -	- virtually no efficiency, less than 7% unexpected responses always require diagnosis

MEASUREMENT ESSENTIALS

2nd Edition

BENJAMIN WRIGHT

MARK STONE

WIDE RANGE, INC.
Wilmington, Delaware