24. GUESSING

What to do about guessing on multiple-choice (MCQ) test items has been a hot problem for 70 years. For some psychometricians the introduction of an extra item parameter for guessing is the way to settle the matter. We review their approach and show how guessing can be better dealt with - detected, diagnosed and managed - by the methods of Rasch measurement.

Webster says "to guess" is to form an opinion from *little or no evidence*. That suggests that when people guess on an MCQ test item, they decide on the basis of little or no ability how to answer. When they are lucky, they guess correctly. It follows that to count lucky guesses as manifestations of ability produces confusion, especially when these counts are combined with correct answers which are the outcome of applied ability. Haphazard combinations of accidental and informative outcomes are bound to be misleading. Webster's also says that "to guess" can mean to arrive at a correct solution by conjecture or intuition. Synonyms for "to guess" include; "to suppose," "to hypothesize." Another alternative definition, "stochastic" from the Greek stokastikos, suggests knowledge arrived at in a probabilistic manner.

Guessing is not done by items, but by persons. When a person, with no knowledge of the correct answer, guesses at random on a multiple-choice question with five alternatives, the probability of success might be as low as $P = .2$. If, however, ability enables the person to eliminate three alternatives as incorrect and hence to reduce the guess to one of two choices, then the probability of success might increase to $P = .5$.

These considerations leave some psychometricians content to deal with guessing as an item parameter. For us, however, the same considerations make clear the ultimate futility of attempting a psychometric solution based on test item characteristics. There are too many personal causes and consequences in guessing for any item guessing parameters to manage.

Guessing can only be addressed and managed by allowing for all of the factors, external and internal, which provoke a person to guess. The most important external factor is the intended use of test results. Internal factors include test administration directions, test format and time allowed. When a passing score allows one to acquire a license to practice a remunerative profession, but a failing score prevents this, a person's approach to a test is different than when the outcome offers no immediate advantage. The uses of test results influence examinee behavior. To expect a psychometric model to resolve these personal influences on the test item or person parameter level is unreasonable.

EXTERNAL FACTORS

Guessing provoked by external factors can only be managed by addressing these factors in their own terms:

1. *Reduce the use of multiple-choice items.* Although this item format enables simplified answer sheet scanning, writers of multiple-choice items seldom overcome the excessive restriction this approach puts upon item construction. Multiple-choice items invite some persons to guess.

2. *Invent better methods of questioning* which eliminate guessing as an active possibility. The use of open-ended questions is one alternative. Providing long, rather than short, lists of possible answers discourages guessing. The versatility and capacity of modern scanners and computers can handle response patterns far more complex than the familiar simple rows of five choices.

3. *Qualify the use of test results* so that they do not force examinees to corrupt their test behavior in order to survive.

4. *Do not administer items that are so hard* that they provoke guessing as the only resort.

5. *Do not make speed a factor in testing.*

MISTAKING GUESSING AS AN ITEM PARAMETER

Psychometricians who deal with guessing as an item parameter argue that better measures result, but is this true? We know that the factors which influence test behavior produce responses to items that consist of idiosyncratic mixtures of ability and guessing. But, if they are combined in some responses but not in others, how can we untangle these components to determine which items have been answered by guessing, and by how much guessing and which have not?

An item guessing parameter assumes that it is the item that causes the guessing and that the effect is the same for every test-taker. Even though some items may sometimes seem to provoke more guessing than others, it is the person, not the item, who initiates guessing, whose momentary state of knowledge and urgency governs the possibility of a lucky guess. Even if some guessing could be handled by an item parameter, a person parameter for guessing behavior would also be needed. We know that some persons guess more than others, a few often, most rarely or never. We also know that no one guesses all of the time.

The item parameter approach to guessing raises the lower end of the item characteristics curve no matter who takes the item.

The asymptotic solution in Figure 24.1 forces a guessing penalty on every person who chooses not to guess. It does this (shaded area in Figure 24.1) by misestimating the item to be easier for non-guessers than it actually is.

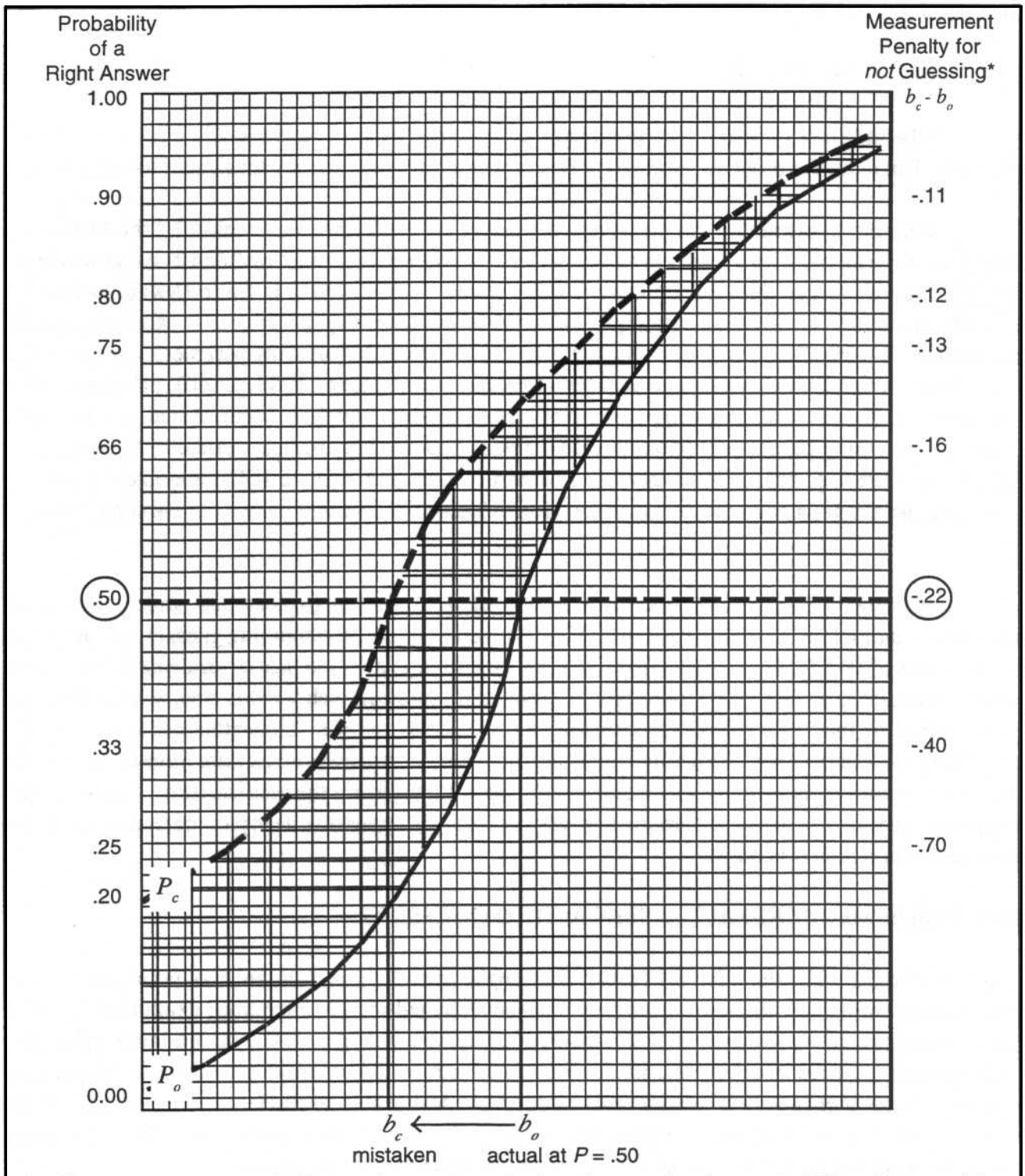
The measurement penalty for not guessing is the distance between b_c and b_o on the measuring variable b in Figure 24.1

$$\left[b_c = d + \log \frac{P_c - C}{1 - P_c} \right] \text{ but } \left[b_o = d + \log \frac{P_o}{1 - P_o} \right]$$

$$b_c - b_o = \log \left[\frac{(P_c - C)}{(1 - P_c)} \frac{(1 - P_o)}{P_o} \right] = \rangle \log \frac{P_o - C}{P_o}$$

when P_o is mistaken in the test score model for P_c .

Figure 24.1
Guessing as a lower asymptote.



* The measurement penalty imposed on persons who do not guess at various probabilities of successful performance.

Guessing, as a lower *asymptote*, modifies the whole curve.

$$P_c = C + (1-C)\frac{\exp(b_c - d)}{1+\exp(b_c - d)}$$ the guesser; $$P_0 = \frac{\exp(b_o - d)}{1+\exp(b_o - d)}$$ the non-guesser

An alternative is to use the lower boundary in Figure 24.2. In this approach there is still a penalty for not guessing, but it is only exacted from non-guessers with performance probabilities below the guessing level C.

ESTIMATION PROBLEMS

If the idea of a guessing item parameter were useful, its application would lead to successful practice. But even the most devoted advocates of a guessing item parameter lament its application.

Attempts to estimate item guessing parameters are uniformly *un*successful, "the likelihood function (of the model with a guessing parameter) may possess several maxima" and its value at infinite ability "may be larger than the maximum value found" when ability is finite (Swaminathan, in Hambleton, 1983, p. 30) and "attempts to estimate the guessing parameter ... are not usually successful" (Hulin, Drasgow & Parsons, 1983, p. 63). "40% of the guessing parameter estimates did not converge even with a sample size of 1593" (Ironson, in Hambleton, 1983, p. 160). "If a test is easy for the group (from which guessing parameters are estimated) and then administered to a less able group, the guessing parameters (from the more able group) may not be appropriate" (Wingersky, in Hambleton, 1983, p. 48). "When dealing with three parameter logistic ICCs, a nonzero guessing parameter *pre*cludes a convenient transformation to linearity" (Hulin, Drasgow & Parsons, 1983, p. 173).

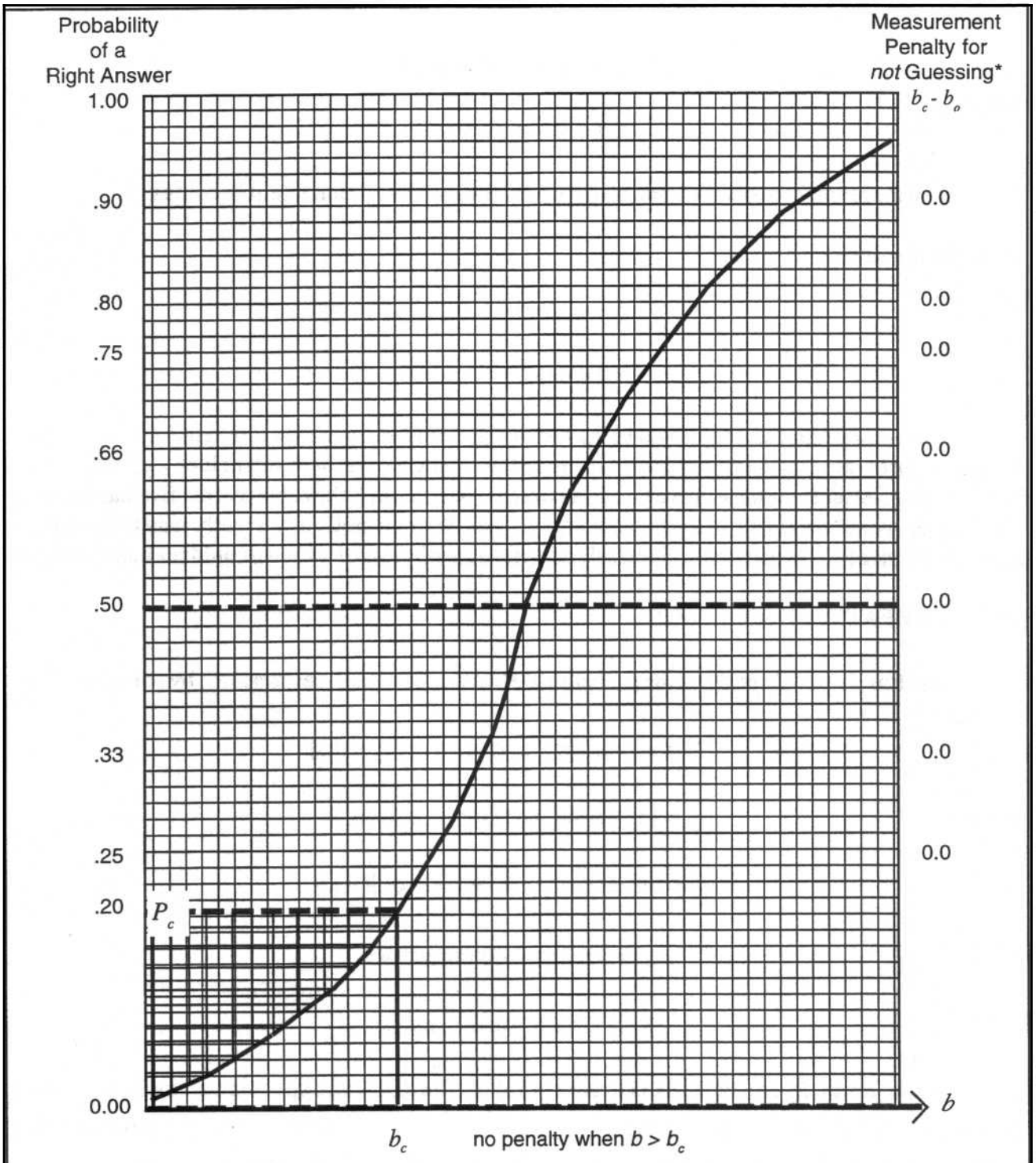
Stocking (1989, p. 41) reports in an extensive study "to explore and understand some apparently anomalous results in various LOGIST-based (a program estimating guessing parameters) applications of IRT that have been obtained from time to time over the past several years" that these same "anomalous results were obtained in simulation studies, such as this one, where data are generated to fit the 3PL (guessing parameter) model" (1989, p. 41). Thus attempts to resolve the guessing problem through estimating a guessing item parameter, *even when data have been created to fit that condition*, have not been successful. Successful practice is the confirmation of theory. The ubiquitous inability to achieve a practical implementation of a guessing item parameter discredits the theories upon which it is based.

THE RASCH MEASUREMENT APPROACH TO GUESSING

To begin with, the external factors that might provoke guessing such as poor test format, abbreviated timing and threatening purpose must be managed so as to encourage examinees to make their responses as uncontaminated as possible by misleading guesses. Maintaining good test management requires constant attention. Failure to reduce the external provocations to guess is sloppy. The problem needs to be addressed by good test design and careful test administration. What must ***not*** be done is to default to a naive presumption that the problem of guessing can be "washed away" by a slick assumption that an item guessing parameter will do the trick.

Guessing is not avoided in Rasch measurement. Guessing is addressed directly by instituting quality control over all response patterns. Consider a score of five on a 10-item test with items positioned in order of increasing difficulty. Both the probabilistic nature of the model and our everyday

Figure 24.2
Guessing as a lower boundary.



* Guessing, as a lower *boundary* modifies the curve only for $P < c$. No penalty above b_c.

$$P_c - P_o = \frac{\exp(b-d)}{1+\exp(b-d)} \text{ for } P > .2$$

$$P_c = .2 \text{ for } P_o < .2$$

experience with typical response patterns lead us to expect patterns like

1 1 1 1 1 0 0 0 0 0 = 5

1 1 1 1 0 1 0 0 0 0 = 5

and, even

1 1 1 0 1 0 1 0 0 0 = 5

The more improbable the pattern, however, the more questionable it becomes.

Consider the pattern

1 1 0 0 0 0 0 1 1 1 = 5

or, worse,

0 0 0 0 0 1 1 1 1 1 = 5.

Our surprise and our objection to the last two patterns are much greater than for the first three. We might speculate that the irregularities in the last two patterns are the result of lucky guessing on the hardest items. After five consecutive wrong answers, it become unbelievable that the five hardest items could be answered correctly on the basis of knowledge. We may not know exactly why this occurred. But we have identified a pattern that is clearly questionable in terms of what we could reasonably expect.

RESPONSE PATTERN ANALYSIS

The Rasch model specifies the probability P_{ni} of dochotomous response x_{ni} by person n to item i to be:

$$P_{ni} = \exp[x_{ni}(b_n - d_i) / [1 + \exp(b_n - d_i)]$$

where b_n = the ability measure of person n

d_i = the difficulty calibration of item i *24.1*

$x_{ni} = 0$ for an incorrect answer

$x_{ni} = 1$ for a correct answer.

Estimates of P_{ni} can be used as expected values for x_{ni}. The expected variance of x_{ni} can be estimated by $[P_{ni}(1 - P_{ni})]$. To estimate a standard residual z_{ni}, we subtract from the observed x_{ni} its expected value P_{ni} and divide by $[P_{ni}(1 - P_{ni})]^{1/2}$ its binomial standard deviation to get

$$z_{ni} = (x_{ni} - P_{ni}) / [P_{ni}(1 - P_{ni})]^{1/2} . \qquad 24.2$$

When the data approximate the measurement model we expect this estimated residual z_{ni} to be distributed symmetrically with a mean of 0, and a variance of 1.

As a rough, but useful, criterion for data fit, we examine the extent to which the distributions of these standard residuals approach

$$z_{ni} \sim N(0,1) \text{ normal}$$
$$z_{ni}^2 \sim \chi_1^2 \text{ chi-square.} \qquad 24.3$$

The reference value 0 for the mean and 1 for the standard deviation and the reference distributions of $N(0,1)$ and χ_1^2 help us to decide whether observed standard residuals deviate unreasonably from model expectations. This examination of residuals helps us to decide whether we can proceed to use these items to make measures and also whether particular persons have failed, at least in part, to respond to the test in a use
ful manner.

When a particular squared residual

$$z_{ni}^2 = (x_{ni} - P_{ni})^2 / P_{ni}(1 - P_{ni}) \qquad 24.4$$

becomes large, we suspect that something unexpected happened when that person n took that item i. A single unexpected response, however, is less indicative of trouble than a pattern of unexpectedly large z_{ni}^2. The accumulated impact of a pattern of large z_{ni}^2 values for a person [or an item] arouses concern for the utility of that person's measure [or that item's calibration].

Consider the responses patterns in Table 24.1.

The circles in Table 24.1 mark unexpected responses. To evaluate the improbability of these responses we replace each instance of an unexpected response by the difference between the ability measure for that person and the difficulty calibration for that item. For Person 1 on Item 4 the unexpected incorrect response associated with person ability $b = -1.2$ and item difficulty $d = -3.9$ produces a difference $(b - d) = (-1.2) - (-3.9) = +2.7$.

This difference 2.7 for Person 1 on Item 4 is placed at the location of that unexpected response in Table 24.2 where we have computed the differences for each instance of an unexpected response circled in Table 24.1.

Unexpected *incorrect* answers have been recorded as $(b - d)$, but unexpected *correct* answers have been recorded as $(d - b)$. We do this because, when the response is incorrect, and $X = 0$, then the index of unexpectedness is $[\exp(b - d)]$, but, when the response is correct, and $X = 1$, then the index becomes $[\exp(d - b)]$.

We record unexpectedness in Table 24.2 as a positive difference, whether from $(b - d)$ or $(d - b)$.

The corresponding values for

$$z^2 = P / (1 - P) = \exp(b - d) \text{ when } X = 0$$
$$\text{and } z^2 = (1 - P) / P = \exp(d - b) \text{ when } X = 1$$

Table 24.1

Some Unexpected Person-to-Item Responses (x)

PERSON	ITEM							NUMBER OF UNEXPECTED RESPONSES	PERSON ABILITY
	4	5	6	7	8	12	14		
1	(0)	1	1	1	1	0	0	1	-1.2
2	1	1	1	(0)	(0)	0	0	2	-1.2
3	1	(0)	1	1	1	0	0	1	-0.6
4	1	1	1	1	1	(1)	0	1	0.0
5	1	1	(0)	(0)	1	(1)	0	3	0.0
6	1	1	(0)	1	(0)	0	(1)	3	0.0
NUMBER OF UNEXPECTED RESPONSES	1	1	2	2	2	2	1	11	
ITEM DIFFICULTY	-3.9	-3.3	-3.3	-2.9	-2.0	1.7	2.8		
"1" = EXPECTED "0" = UNEXPECTED SINCE THESE PERSONS ARE *ABOVE* -2.0 IN ABILITY						"0" = EXPECTED "1" = UNEXPECTED SINCE THESE PERSONS ARE *BELOW* +1.7 IN ABILITY			

can then be evaluated for the improbability of the response. These z^2 values, which are taken from Column 2 of Table 24.4 (Best Test Design, Wright & Stone, 1979, p. 73) have been entered in Table 24.3.

Table 24.4 gives values of $z^2 = \exp(b-d)$ for unexpected incorrect answers $x = 0$ or values of $z^2 = \exp(b-d)$ for unexpected correct answers $x = 1$.

The entry C_x in Column 1 of Table 24.4 is $C_0 = (b-d)$ when the response is incorrect and $x = 0$ and $C_1 = (d-b)$ when the response is correct and $x = 1$.

We locate the difference +2.7 for the $|b-d|$ of Person 11 on Item 4 in Column 1 of Table 24.4 and read the corresponding z^2 in Column 2 as 15. This value and all of the other values for the differences in Table 24.2 have been recorded in Table 24.3 which now contains the z^2 for each instance of unexpectedness that we have observed for the six persons and seven items. The margins of Table 24.3

Table 24.2

Differences *(b-d)* Between Person Ability and

Item Difficulty for Unexpected Responses

PERSON	ITEM							PERSON ABILITY
	4	5	6	7	8	12	14	
1	2.7							-1.2
2				1.7	0.8			-1.2
3		2.7						-0.6
4						1.7		0.0
5			3.3	2.9		1.7		0.0
6			3.3		2.0		2.8	0.0
ITEM DIFFICULTY	-3.9	-3.3	-3.3	-2.9	-2.0	1.7	2.8	
	"1" = EXPECTED "0" = UNEXPECTED SINCE THESE PERSONS ARE *ABOVE* -2.0 IN ABILITY					"0" = EXPECTED "1" = UNEXPECTED SINCE THESE PERSONS ARE *BELOW* +1.7 IN ABILITY		

give the sums of these z^2 for each person and each item. These sums indicate how unexpected the *patterns* of person or item responses are.

Column 3 of Table 24.4 shows $p = 1/(1+z^2)$, the model improbability of each observed response. This value provides a significance level for the null hypothesis of acceptable fit for any particular response. With our example of $(b - d) = 2.7$ we find a significance level of .06 in the table, against the null hypothesis that the response of Person 1 to Item 4 is according to the model.

QUALITY CONTROL

The evaluation of response patterns is a quality control procedure. In Rasch measurement, quality control over response patterns is implemented by determining the fit of response patterns to modeled expectations. Fit, or response plausibility, is determined from the difference between the estimates of person ability b and item difficulty d for each person/item contact. When this difference is positive, the item should be easy for the person. The more positive the difference, the easier the item and hence the greater our expectation that the person will succeed. Similarly, as the difference between

Table 24.3

Fit Mean Squares (z^2) for Unexpected Responses

PERSON	ITEM							PERSON MISFIT TOTAL
	4	5	6	7	8	12	14	
1	15							15
2				6	2			8
3		15						15
4						6		6
5			(27)	18		6		(51)
6			(27)		7		17	(51)
ITEM MISFIT TOTAL	15	15	(54)	24	9	12	17	146
	"1" = EXPECTED "0" = UNEXPECTED SINCE THESE PERSONS ARE *ABOVE* -2.0 IN ABILITY					"0" = EXPECTED "1" = UNEXPECTED SINCE THESE PERSONS ARE *BELOW* +1.7 IN ABILITY		

person ability and item difficulty becomes more negative, the item should be more difficult for the person, and our expectation of failure increases.

The response pattern produced by each person is evaluated for the amount of misfit occurring. The diagnosis of patterns is expedited by plotting to show each pattern's shape and by summarizing the misfit in that particular pattern. A summary fit statistic is computed for each person and each item.

Figure 24.3 shows a response pattern that suggests guessing with an initial ability measure of $b = 3.2$. Four easy items were answered correctly followed by five items of increasing difficulty answered incorrectly followed finally by *two quite difficult items answered correctly*. Our attention is attracted to these last two most difficult items with correct responses following five easier items answered incorrectly. These last two correct responses are implausible.

Table 24.5 shows the residual analysis of the original pattern of responses and of the corrected pattern. We can compute two ability estimates for this person. One, at $b = 3.2$, is based upon the original full pattern. The other, much lower, at $b' = 1.7$, is based on deleting the last two implausible items. We question whether the original ability estimate $b = 3.2$ is a good indicator of this person's position on the variable because the response pattern misfit is $t = 5.3$. The corrected pattern fit of $t' = -1.2$ is more

Table 24.4

Some Misfit Statistics

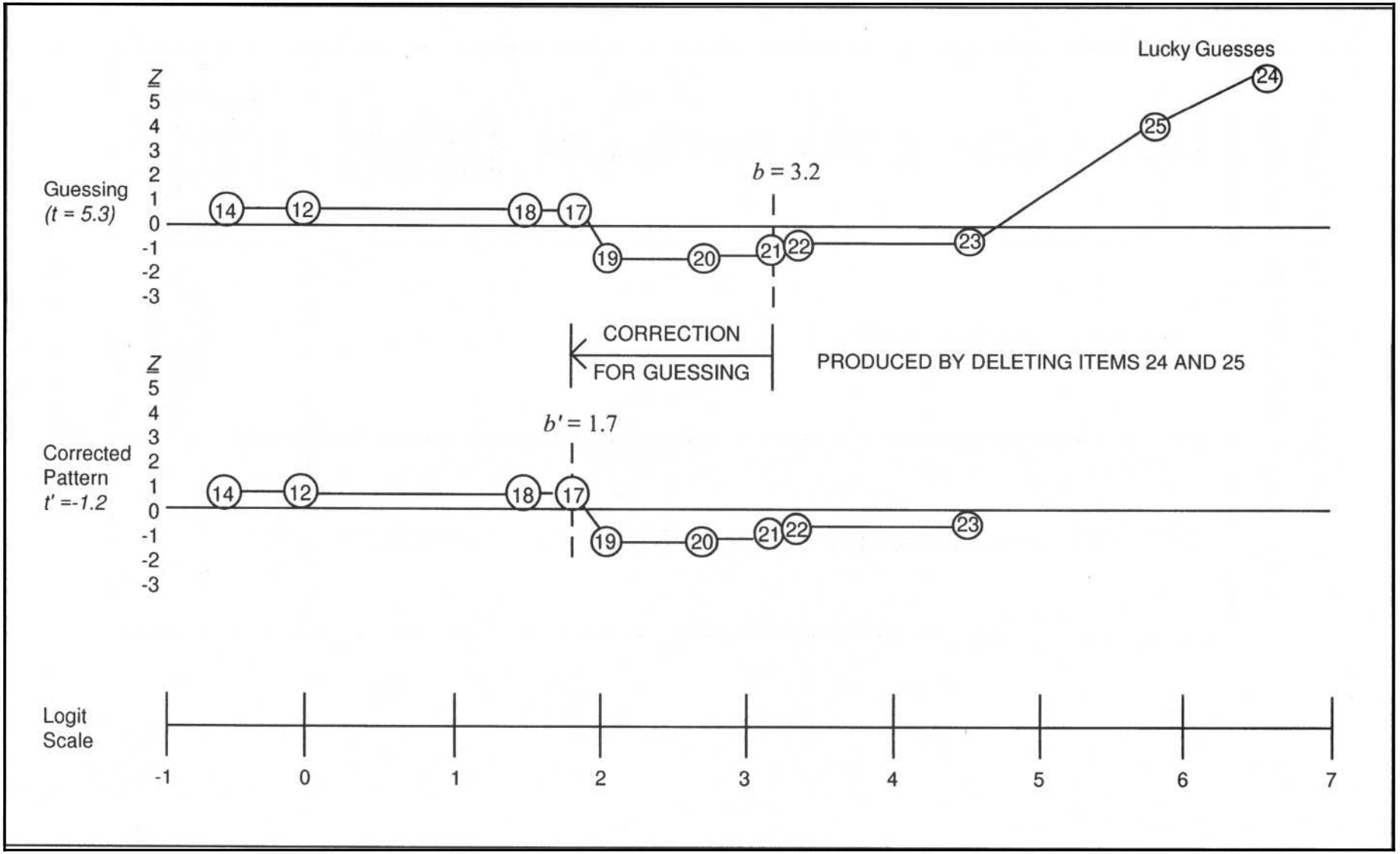
1. DIFFERENCE BETWEEN PERSON ABILITY AND ITEM DIFFICULTY C_x*	2. SQUARED STANDARDIZED RESIDUAL $z^2=expC$	3. IMPROBABILITY OF THE RESPONSE $P=1/(1+z^2)$
-0.6, 0.4	1	.50
0.5, 0.9	2	.33
1.0, 1.2	3	.25
1.3, 1.5	4	.20
1.6, 1.7	5	.17
1.8, 1.8	6	.14
1.9, 2.0	7	.12
2.1	8	.11
2.2	9	.10
2.3	10	.09
2.4	11	.08
2.5	12	.08
2.6	13	.07
2.7	15	.06
2.8	16	.06
2.9	18	.05
3.0	20	.05
3.1	22	.04
3.2	25	.04
3.3	27	.04
3.4	30	.03
3.5	33	.03
3.6	37	.03
3.7	40	.02
3.8	45	.02
3.9	49	.02
4.0	55	.02
4.1	60	.02
4.2	67	.02
4.3	74	.01
4.4	81	.01
4.5	90	.01
4.6	99	.01

* For incorrect responses when $x = 0$ then $C_o=(b\text{-}d)$.

For correct responses when $x = 1$ then $C_1=(d\text{-}b)$.

Figure 24.3

Correcting a guessing pattern.



acceptable. Which estimate we decide is more useful depends upon what we think about the responses of the person to these two items. If we think that these responses are implausible, that it is unlikely that this person would get these last two items correct after five failures, then we might take the corrected $b' = 1.7$ as the more useful estimate of this person's measure.

Statistical analysis alone cannot tell which estimate is more useful, but it can detect and arrange the available information into a concise and objective summary to use as part of our evaluation of the person. Persons who guess on multiple choice items may succeed on difficult items more often than their abilities predict. This could make them appear more able, especially if many items are too difficult for them. This is because their frequency of success would not decrease as item difficulty increased. A similar but opposite effect occurs when able persons become careless with easy items, making these persons appear less able.

Item responses affected by guessing express the simultaneous influence of more than one variable. There is the ability to be measured and, in addition, there is the tendency to guess. The "guessingness" of the item may or may not be a simple function of its difficulty on the main variable, or, if a multiple choice item, of its distractors. For the person being measured, at least, two quite different variables are involved. One is ability, the other is inclination to guess. The accurate measurement of either variable is threatened by the active presence of the other. In our empirical experience, when guessing does occur, it is dominated by the specific individuals who do the guessing and not by particular items, unless the items are poorly constructed.

When we detect a significant misfit in a response record, diagnose the response pattern and identify possible reasons for its occurrence, it is finally necessary to decide whether an improved measure can or should be determined. Whether a statistically "corrected" measure is "fair" for the person or "proper" for the testing authority cannot be settled by statistics. Nevertheless, knowing how a measure can be corrected objectively gives us a better understanding of the possible meaning in a person's performance.

For other examples of misfit patterns see Chapter 17 (p.143), Information and Misfit Analysis and Best Test Design (Wright & Stone, 1979, pages 165-190).

TAILORED MEASURING

In situations where we think that guessing may be influenced by test format as, for example, when we think a person may guess at random over m multiple-choice alternatives, we can use the guessing probability of $1/m$ as a threshold below which we suppose guessing might occur, as in Figure 24.2. To guard our measures against this kind of guessing we can delete all items from a person response record which have difficulty greater than $b + \log(m - 1)$ where b is the person's initial estimated ability. After these deletions we reestimate the person's ability from the remaining items attempted. When we do this, we are taking the position that when items are so difficult that a person can do better by random guessing than by actually trying, then, whatever the person's responses may be, such items should not be used to estimate the person's ability.

The procedure is:

1) When several unexpected responses are "correct" beyond some set fit statistic, say $t > 3$, suggesting the possibility of lucky guessing on the part of this particular person, delete all

Table 24.5

Correcting a Guessing Pattern

CASE DESCRIPTION	RESPONSE STATISTIC	ITEM NAME AND DIFFICULTY (IN DIFFICULTY ORDER)										
	ITEM NAME:	#14	#12	#18	#17	#19	#20	#21	#22	#23	#25	#24
	ITEM DIFFICULTY:	-0.5	-0.1	1.4	1.9	2.0	2.9	3.3	3.3	4.5	5.8	6.3
		RESPONSE PATTERN										
							b = 3.2				DELETE	
"Guessing" Pattern (b = 3.2)	x	1	1	1	1	0	0	0	0	0	1	1
	$(2x\text{-}1)(d\text{-}b)$	-3.7	-3.3	-1.8	1.3	1.2	0.3	-0.1	-0.1	-1.3	2.6	3.1
	z^2	0.0	0.0	0.2	0.3	3.3	1.4	0.9	0.9	0.3	13.5	22.2
	z	0.2	0.2	0.4	0.5	-1.8	-1.2	-1.0	-1.0	-0.5	3.7	4.7
					CORRECTION FOR GUESSING *							
				b' = 1.7								
Corrected Pattern (b' = 1.7)	x	1	1	1	1	0	0	0	0	0		
	$(2x\text{-}1)(d\text{-}b')$	-2.2	-1.8	-0.3	-0.2	-0.3	-1.2	-1.6	-1.6	-2.8		
	z^2	0.1	0.2	0.7	1.2	0.7	0.3	0.2	0.2	0.1		
	z	0.3	0.4	0.9	1.1	-0.9	-0.5	-0.4	-0.4	-0.2		

* *"Guessing" correction rule*: for m-choice items delede $d > [b + ln\,(m - 1)]$.
i.e., if $m = 5$, and $b = 3.2$, then delete any items with $d > 3.2\ ln\,(4) = 3.2 + 1.4 = 4.6$, i.e., items #25 and #24.

	MEASUREMENT						RESIDUAL ANALYSIS		
	Score r	Relative Score $f = r/L$	Relative Ability X_{fw}	Error Coefficient $C_{fw}^{1/2}$	Ability b	Error s	Sum of Squares $\sum z^2$	Mean Square v	Fit Statistic t
"Guessing" Pattern (b = 3.2)	6	.55	0.4	2.9	3.2	0.9	43.0	4.3	5.3 misfit
Corrected Pattern (b' = 1.7)	4	.44	-0.4	2.4	1.7	0.8	3.7	0.5	-1.2

Correction in measure: 1.7 - 3.2 = 1.5.

$v = \sum z^2/(L-1)$

$t = [\ln(v) + (v-1)][(L-1)/8]^{1/2}$

"too hard" items from this particular response pattern, that is, all items with $d > [b + \log (m - 1)]$ where m is the number of alternatives.

2) Compute a new ability estimate after the deletion of the too hard items and make another analysis of fit.

Steps 1 and 2 can be reiterated until successive values of b become stable and fit becomes acceptable. When this procedure is applied to response patterns generated entirely by guessing, the illegitimate responses are peeled away one at a time until the entire pattern is gone.

The use of a quality control process through misfit analysis of response patterns is the Rasch measurement way of dealing with guessing. In Rasch measurement we do not accept guessing as something to be tolerated when it can be avoided by external means, nor do we leave guessing to faulty estimation procedures produced by unworkable models. Instead, we arrange the testing experience so that guessing is least likely to occur and then use the quality control procedure of fit analysis to monitor all response patterns for any manifestations of whatever lucky guessing may occur.

MEASUREMENT ESSENTIALS

2nd Edition

BENJAMIN WRIGHT

MARK STONE

WIDE RANGE, INC.
Wilmington, Delaware