A Cautionary Note on Estimating the Reliability of a Mastery Test with the Beta-Binomial Model

Rand R. Wilcox
University of Southern California

Based on recently published papers, it might be tempting to routinely apply the beta-binomial model to obtain a single administration estimate of the reliability of a mastery test. Using real data, the paper illustrates two practical problems with estimating reliability in this manner. The first is that the model might give a poor fit to data, which can seriously affect the reliability estimate, and the second is that inadmissible estimates of the parameters in the beta-binomial model might be obtained. Two possible solutions are described and illustrated.

In recent years, efforts have been directed toward deriving ways of studying and characterizing mastery and criterion-referenced tests. A summary of the statistical and psychometric techniques that have evolved can be found in the 1980 special issue of *Applied Psychological Measurement* (see, also, Hambleton, Swaminathan, Algina, & Coulson, 1978). One approach that has received considerable attention can be described as follows: Suppose two randomly parallel test forms both consist of n dichotomously scored items. For a randomly sampled examinee, let x and y be the observed scores on the two test forms and let $f(x, y)$ be the joint probability function of x and y for the population of examinees. If the same passing score, say x_0, is used on both test forms, the proportion of agreement is defined to be

$$P = \sum_{x=x_0}^{n} \sum_{y=x_0}^{n} f(x,y) + \sum_{x=0}^{x_0-1} \sum_{y=0}^{x_0-1} f(x,y). \qquad [1]$$

Many other methods have been proposed for characterizing mastery tests, but at a minimum it is desired that P be reasonably close to one (see, e.g., Traub & Rowley, 1980).

Frequently, it is difficult to administer two randomly parallel tests to a random sample of examinees. Accordingly, efforts have been made to derive an estimate of P based on the observed scores of only one test form. A general approach to this problem is as follows: For a specific examinee, assume that the probability of an observed score x is $f(x \mid \theta)$, where θ is some unknown parameter, possibly vector valued. For the randomly parallel test, let $f(y \mid \theta)$ be the probability of an observed y;

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 5, No. 4, Fall 1981, pp. 531–537

0146-6216/81/040531-07$1.35

and suppose $f(x \mid \theta)$ and $f(y \mid \theta)$ are independent and that they have the same parametric form. If $g(\theta)$ is the density function of θ over the population of examinees, then

$$f(x,y) = \int f(x|\theta)f(y|\theta)g(\theta)d\theta. \qquad [2]$$

Once a specific form for $f(x \mid \theta)$ and $g(\theta)$ is assumed, it is frequently possible to estimate $g(\theta)$, which yields an estimate of $f(x, y)$. This, in turn, yields an estimate of P via Equation 1.

In the statistical literature the single administration estimate of P described above is known as an empirical Bayes approach to prediction analysis. (For general results on prediction analysis, see Aitchison and Dunsmore, 1975.)

Huynh (1976) has given a detailed account of how to estimate P for the special case where $f(x \mid \theta)$ (and $f(y \mid \theta)$) are assumed to be binomial, and where $g(\theta)$ is assumed to belong to the beta family of distributions. Note, however, that Huynh concentrates on estimating Cohen's (1960) kappa rather than P once the estimate of $f(x, y)$ is available (cf. Divgi, 1980). Since Huynh's paper, several investigations of the beta-binomial model have been reported that are relevant to estimating reliability via Equation 2. For example, Subkoviak (1978) compared it to three other estimates of P and concluded that all four methods gave good results but that the beta-binomial model seemed to be the best for general use.[1]

Based on the studies cited above, it might be tempting to routinely apply the beta-binomial model when estimating the proportion of agreement or some related coefficient such as Cohen's kappa. In practice, though, there are at least two practical problems that might arise. First, the beta-binomial model might give a poor fit to the data (Keats, 1964), which, as illustrated below, might affect the estimate of P. Second, an estimate of the parameters in the beta-binomial model might not exist. That is, all five estimation procedures described by Wilcox (1979), including the maximum likelihood estimate proposed by Griffiths (1973), might yield negative values even though the model assumes they are positive. Negative estimates can occur even when the model holds, or they might occur because the model is completely inappropriate.

Using real test data, this paper illustrates that the beta-binomial model should not be automatically assumed to give a good estimate of P. Instead, an investigator should first verify that the model gives a reasonable fit to the data. Experience suggests that the beta-binomial model will frequently give a good fit; but when it fails to do so, some other estimate of P should be used. The paper also describes and illustrates two alternative estimates that might be used when a poor fit with the beta-binomial model is obtained.

Two Alternatives to the Beta-Binomial Model

Temporarily consider a single examinee responding to n dichotomously scored items. The binomial error model assumes that

$$f(x|\theta) = \binom{n}{x} \theta^x(1-\theta)^{n-x} \qquad [3]$$

This assumption is theoretically justified when items are randomly sampled from an infinite item pool (or a finite pool with replacement), when the examinee's responses are independent from one another, and when the probability of a correct response is θ for every randomly sampled item. In

[1] Additional empirical support for the beta-binomial model can be found in Gross and Shulman (1980). For further results and comments on P, see Algina and Noe (1978), Huynh (1979), Divgi (1980), Traub and Rowely (1980), and Subkoviak (1980). For a recent review of the beta-binomial model, see Wilcox (in press).

many instances items are not randomly sampled; and even when they are, it is customary for every examinee to respond to the same n items. Thus, it is not surprising to find situations where Equation 3 gives unsatisfactory results.

When trying to find a probability function that gives a good fit to data, probably three of the best known and most frequently employed distributions are the binomial, Poisson, and negative-binomial (Johnson & Kotz, 1969). Thus, when the beta-binomial model is unsatisfactory, it is reasonable to consider replacing Equation 3 with a Poisson or negative-binomial distribution. Of course, the Poisson distribution is not new to psychometric theory (Lord & Novick, 1968, chap. 21), but it has not been used in the manner outlined below. The important aspect of the Poisson distribution is that it frequently gives a good approximation to an observed distribution associated with an infrequently occurring event (e.g., Johnson & Kotz, 1969). This suggests that if observed test scores are skewed, and if the beta-binomial does not give a good fit to data, then a Poisson model might be considered. This is exactly what was done in Wilcox (in press) using real test data.

The Gamma-Poisson Model

Let $w = n - x$ and $z = n - y$ be the number of incorrect responses given by an examinee on the first and second test forms, respectively. Begin by replacing Equation 3 with the assumption that the probability function of w, as well as z, is Poisson with parameter η. Symbolically

$$f(w|\eta) = e^{-\eta}\eta^{w}/w! \qquad [4]$$

The reason for working with w and z, rather than x and y, is that the data in the example is skewed to the right. If the observed frequencies had been skewed to the left, x and y would have been used (see Wilcox, in press).

Assume also that for the population of examinees, η has a gamma distribution. The motivation for this assumption is that it is typically made for the Poisson case, it is mathematically convenient, and it has given good results with mental test data (Wilcox, in press). If $f(w \mid \eta)$ and $f(z \mid \eta)$ are assumed to be independent, results in Aitchison and Dunsmore (1975) indicate immediately that

$$f(w) = \frac{\Gamma(\alpha+w)}{\Gamma(\alpha)\Gamma(w+1)} \left(\frac{\beta}{\beta+1}\right)^{w} \left(\frac{1}{\beta+1}\right)^{\alpha} \qquad [5]$$

i.e., the marginal probability function of w is negative binomial. The parameters α and β can be estimated as follows: Let $\bar{w}$ and s^2 be the sample mean and variance of w for a random sample of examinees. Then $\hat{\beta} = (s^2/\bar{w})-1$ and $\hat{\alpha} = \bar{w}/\hat{\beta}$ estimate β and α, respectively. Three other estimates of α and β are also available (Johnson & Kotz, 1969).

Again, referring to Aitchison and Dunsmore (1975),

$$f(z|w) = \frac{\Gamma(\alpha+w+z)}{\Gamma(\alpha+w)\Gamma(z+1)} \left(\frac{\beta}{2\beta+1}\right)^{z} \left(\frac{\beta+1}{2\beta+1}\right)^{\alpha+w} \qquad [6]$$

Since $f(w, z) = f(w)f(z \mid w)$, there is an estimate of P once α and β are determined.

The Gamma Product-Ratio Poisson Model

The other model considered also assumes Equation 4, but η is assumed to have a "gamma product-ratio" distribution (Sibuya, 1979). In this case

$$f(w) = \frac{\Gamma(w+\alpha)\Gamma(\beta+\gamma)\Gamma(w+\beta)\Gamma(\alpha+\gamma)}{\Gamma(w+1)\Gamma(\alpha)\Gamma(\beta)\Gamma(\gamma)\Gamma(\alpha+\beta+\gamma+w)} \quad [7]$$

where $\alpha, \beta, \gamma > 0$ are unknown parameters. Note that two alternative names for Equation 7 are generalized Waring and negative-binomial beta. Also, the parameters α and β in Equation 7 are different from those in Equation 6.

To estimate α and β, and γ, first note that the first three factorial moments are

$$\mu_1 = \alpha\beta/(\gamma-1) \quad [8]$$

$$\mu_2 = \alpha(\alpha+1)\beta(\beta+1)/[(\gamma-1)(\gamma-2)] \quad [9]$$

$$\mu_3 = \alpha(\alpha+1)(\alpha+2)\beta(\beta+1)(\beta+2)/[(\gamma-1)(\gamma-2)(\gamma-3)] \quad [10]$$

It follows that

$$\left(\frac{\mu_2}{\mu_1} - \mu_1\right)\gamma - \alpha - \beta = \frac{2\mu_2}{\mu_1} - \mu_1 + 1 \quad [11]$$

and

$$\left(\frac{\mu_3}{\mu_2} - \mu_1\right)\gamma - 2\alpha - 2\beta = \frac{3\mu_3}{\mu_2} - \mu_1 + 4 \quad [12]$$

Thus, if $\hat{\mu}_i$ is the usual estimate of μ_i ($i = 1, 2, 3$), there is an estimate of γ, say $\hat{\gamma}$. Substituting $\hat{\gamma}$ and $\hat{\mu}_1$ and $\hat{\mu}_2$ into Equations 8 and 9 yields

$$\alpha = \hat{\mu}_1(\hat{\gamma}-1)\beta^{-1} \quad [13]$$

$$\alpha+\beta = \frac{\hat{\mu}_2}{\hat{\mu}_1}(\hat{\gamma}-2)-\hat{\mu}_1(\hat{\gamma}-1)-1 \quad [14]$$

Substituting the right-hand side of Equation 13 for α in Equation 14 yields a quadratic equation for β. In terms of the marginal density in Equation 7, either estimate of β can be used, since the other estimate of β will correspond to α and since Equation 7 is symmetric in α and β.

Finally, to estimate P with Equation 1, note that

$$f(w,z) = \frac{\Gamma(\alpha+w)\Gamma(\alpha+z)\Gamma(\beta+\gamma)\Gamma(\beta+w+z)\Gamma(2\alpha+\gamma)}{\Gamma(\alpha)\Gamma(\alpha)\Gamma(w+1)\Gamma(z+1)\Gamma(\beta)\Gamma(\gamma)\Gamma(2\alpha+\beta+\gamma+z+w)} \quad [15]$$

One way to establish this result is to assume $f(w \mid \theta)$ is negative-binomial and that $g(\theta)$ is beta (which is equivalent to assuming Equation 7) and then performing the integration in Equation 2.

Numerical Illustrations

This section uses real data to illustrate the practical advantages of estimating P with the two alternative estimates described above.

First, consider the data reported in Keats (1964). Keats has indicated that the beta-binomial model gives a poor fit to the observed test scores, but as noted in Wilcox (in press), the gamma-Poisson model gives a reasonably good fit. The test had $n = 30$ items, and Keats has reported observed test scores for 1,000 examinees. If P is estimated with the beta-binomial model, the result is .90. If the gamma-Poisson model is used, the estimate is .81. The third estimate of P does not apply, since the estimates of the parameters in Equation 15 are inadmissible. Note that the reliability estimates used by Subkoviak (1976), as well as by Marshall and Haertel (1975), also assume that the binomial error model holds. Since the beta-binomial model gives a poor fit to data, there is some doubt about whether these estimates should even be considered.

As another illustration, suppose there is an $n = 15$ item test with a passing score of $x_0 = 10$. Further, suppose there are test scores as reported in Table 1. These results are based on real data reported in Irwin (1968), but they do not represent test scores. The point is that observed frequencies that are skewed, as are the frequencies in Table 1, might be obtained, in which case it might be better, or even necessary, to replace the beta-binomial model with something else.

Table 1
Observed Error on an n=15 Item Test

W	Observed Frequency		Expected Frequencies for Gamma Product-Ratio Poisson
0	239		240
1	98		105
2	57		49
3	33		24
4	9		12
5	2		7
6	2		4
7	1		
8	0		
9	4		
10	1		
11	0	7	6
12	0		
13	1		
14	0		
15	0		

Note: The significance level of the chi-square goodness of fit test was .03.

For the data in Table 1, the estimates of the parameters in the beta-binomial model are negative, and so an estimate of P cannot be made. Suppose instead Equation 7 holds. It follows that $\hat{\alpha} = 5.2162$, $\hat{\beta} = 1.297$, and $\hat{\gamma} = 7.7967$. Thus, the estimate of P is .97. If, instead, the gamma-Poisson model is used, the estimate of P is again .97.

Conclusion

It is not being suggested that the gamma-Poisson or the gamma product-ratio Poisson should replace the beta-binomial model, or that it is more flexible than the beta-binomial model when trying to find a model that gives a good fit to data. It is being argued that the beta-binomial model should not be assumed to give a good fit to data. Moreover, if a poor fit is obtained, an attempt should be made to find a model that gives a reasonable fit before the proportion of agreement is estimated.

It was suggested that the gamma-Poisson or the gamma product ratio Poisson be considered when the beta-binomial model gives a poor fit to data, particularly when the observed sample distribution is skewed. As was illustrated, this can lead to a substantially different estimate of P.

It would be convenient to have a more flexible model when trying to get a good fit to data which in turn could be used to estimate P. There are models more flexible than the beta-binomial model in terms of getting a good fit to data, for example, the generalized hypergeometric model (Keats, 1964), but these models do not yield an estimate of P because they do not provide an estimate of the true score distribution that can be used in Equation 2.

It was mentioned that existing methods of estimating the parameters of the beta-binomial model might yield inadmissible (negative) values, even when the model holds. If the model does hold, the best way to avoid this problem is to use a reasonably large number of examinees. If the problem still occurs, it *might* be because the beta-binomial model is inappropriate, in which case one of the alternative estimates outlined above might be used. However, if the model does indeed hold, the problem of inadmissible estimates is still unresolved.

Finally, if dichotomously scored items are randomly selected and scored sequentially until the examinee gets, say, M items correct, the negative binomial is the appropriate distribution for a single examinee. This particular distribution is not included in Aitchison and Dunsmore (1975); with the results reported here, however, an estimate of P can be obtained.

References

Aitchison, J., & Dunsmore, I. R. *Statistical prediction analysis.* London: Cambridge University Press, 1975.

Algina, J., & Noe, M. J. A study of the accuracy of Subkoviak's single administration estimate of the coefficient of agreement using two true-score estimates. *Journal of Educational Measurement,* 1978, *15,* 101-110.

Cohen, J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement,* 1960, *20,* 37-46.

Divgi, D. R. Group dependence of some reliability indices for mastery tests. *Applied Psychological Measurement,* 1980, *4,* 213-218.

Griffiths, D. A. Maximum likelihood estimation for the beta-binomial distribution and an application to the household distribution of the total number of cases of a disease. *Biometrics,* 1973, *29,* 637-648.

Gross, A. L., & Shulman, V. The applicability of the beta-binomial for criterion-referenced testing. *Journal of Educational Measurement,* 1980, *17,* 195-202.

Hambleton, R. K., Swaminathan, H., Algina, J., & Coulson, D. Criterion-referenced testing and measurement: A review of technical issues and developments. *Review of Educational Research,* 1978, *48,* 1-47.

Huynh, H. On the reliability of decisions in domain-referenced testing. *Journal of Educational Measurement,* 1976, *13,* 253-264.

Huynh, H. Statistical inference for two reliability indices in mastery testing based on the beta-binomial model. *Journal of Educational Statistics,* 1979, *4,* 231-246.

Irwin, J. O. The generalized Waring distribution applied to accident data. *Journal of the Royal Statistical Society, Series A,* 1968, *131,* 205-225.

Johnson, N., & Kotz, S. *Discrete distributions.* New York: Wiley, 1969.

Keats, J. A. Some generalizations of a theoretical distribution of mental test scores. *Psychometrika,* 1964, *29,* 215-231.

Lord, F. M., & Novick, M. R. *Statistical theories of mental test scores.* Reading MA: Addison-Wesley, 1968.

Marshall, J. L., & Haertel, E. H. *A single-administration reliability index for criterion-referenced tests: The mean split-half coefficient of agreement.* Paper presented at the annual meeting of the American Educational Research Association, April 1975.

Sibuya, M. Generalized hypergeometric, digamma, and trigamma distributions. *Annals of the Institute of Statistical Mathematics,* 1979, *31,* 373-390.

Subkoviak, M. J. Estimating reliability from a single administration of a criterion-referenced test. *Journal of Educational Measurement,* 1976, *13,* 265-276.

Subkoviak, M. Decision-consistency approaches. In R. Berk (Ed.), Criterion-referenced measurement: The state of the art. Baltimore: The Johns Hopkins University Press, 1980.

Traub, R., & Rowley, G. L. Reliability of test scores and decisions. *Applied Psychological Measurement,* 1980, *4,* 517-545.

Wilcox, R. R. Estimating the parameters of the beta-binomial distribution. *Educational and Psychological Measurement,* 1979, *31,* 527-535.

Wilcox, R. R. A review of the beta-binomial model and its extensions. *Journal of Educational Statistics,* in press.

Acknowledgments

The project presented or reported herein was performed pursuant to a grant from the National Institute of Education, Department of Health, Education, and Welfare. However, the opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education, and no official endorsement by the National Institute of Education should be inferred. A portion of this work was completed at the Center for the Study of Evaluation, University of California, Los Angeles.

Author's Address

Send requests for reprints or further information to Rand R. Wilcox, Department of Psychology, University of Southern California, Los Angeles CA 90007.