A NOTE ON THE STANDARDIZATION OF A FIT STATISTIC CONDITIONAL ON TEST SCORE

D.N.M. de Gruijter*

Abstract

In the last decade many person fit statistics or appropriateness indices have been suggested. A recent proposal by Molenaar and Hoijtink concerns a fit statistic for the Rasch model. Molenaar and Hoijtink obtain approximate conditional distributions using the method of moments. For large tests other approaches are needed. Kogut resorted to simulations. In the present study the method of moments is used in connection with an intermediate distribution.

Introduction

Some examinees who take a test, may have an aberrant score pattern, a pattern which is very improbable considering the other score patterns. Such deviant patterns may arise by chance. They may, however, also arise from unwanted factors like cheating. For this reason it is important to find ways to measure deviance or fit of response patterns. In the last decade many proposals for person fit or appropriateness indices have been made. Several of these indices have been discussed and compared in studies by Harnisch & Linn (1981), Harnisch & Tatsuoka (1983), Rudner (1983), Drasgow, Levine & McLaughlin (1987), and Molenaar & Hoijtink (1989). Fit statistics proposed by Wright & Stone (1979), and Levine & Rubin (1979), Drasgow, Levine & Williams (1985), Drasgow, Levine & McLaughlin (1987), and Molenaar & Hoijtink (1989), are based on IRT. Other indices are Sato's caution index (Sato, 1975), Van der Flier's deviance index (1977), and a modified caution index (Harnisch & Linn, 1981).

Drasgow et al. (1987) argue that an index should satisfy two criteria. First, an index should be standardized. An index distribution should not depend on ability level: there should be no confounding between ability level and typicality. Drasgow et al. suggest the possibility to obtain the distribution of an index conditional on ability level: this can be done in principle for all indices. They think, however, that the computation of conditional distributions is too time consuming to be of practical relevance. This is true for exact computation with long tests, but Kogut's work (1987) demonstrates the feasibility of the estimation of a distribution through sampling.

* Bureau Onderzoek van Onderwijs, Boerhaavelaan 2, 2334 EN Leiden, 071-277170

The second criterion is relative power: given a particular rate of classifying response patterns of a normal group as deviant, which index has the highest rate of identifying patterns resulting from deviant response processes, correctly?

Drasgow, Levine & Williams (1985) suggested standardized indices for dichotomous and polytomous items based on IRT. Their index for dichotomous items can be written as

$$z = [\ell_0 - M(\hat{\theta})]/s(\hat{\theta}). \tag{1}$$

In this formula ℓ_0 is the logarithm of the likelihood evaluated at the ML-estimate $\hat{\theta}$. Let us take a test with n items. With estimated proportion correct $P_i(\hat{\theta})$ for item i and with $Q_i(\hat{\theta})=1-P_i(\hat{\theta})$ mean and variance of the log likelihood conditional on $\theta=\hat{\theta}$ are approximated as

$$M(\hat{\theta}) = \sum_{i=1}^{n} [P_i(\hat{\theta})\log P_i(\hat{\theta})+Q_i(\hat{\theta})\log Q_i(\hat{\theta})] \tag{2}$$

and

$$s^2(\hat{\theta}) = \sum_{i=1}^{n} P_i(\hat{\theta})Q_i(\hat{\theta})\{\log[P_i(\hat{\theta})/Q_i(\hat{\theta})]\}^2. \tag{3}$$

Drasgow et al. (1985) suggest that z is approximately normal. Molenaar and Hoijtink (1989) have raised two objections against z. First, they demonstrated that the normality assumption is not warranted, at least for small and medium sized tests. Secondly, they noted a conceptual problem with the choice of standardization in Equation 1. Equation 1 is based on the assumption that all response patterns are possible. For most of these response patterns estimated ability would be different from the obtained estimate $\hat{\theta}$ on which the value of z is based. Another way of standardizing z might be more appropriate.

Molenaar and Hoijtink suggested to condition the fit statistic on total score t. They did so in the context of the Rasch model in which total score is a sufficient statistic for $\hat{\theta}$. In this contribution it is argued that conditoning on total score might be the most appropriate thing to do for other IRT-models as well. In the next section the Molenaar statistic is introduced, and the computational problems are illuminated. Subsequently, a new computational procedure is discussed.

Fit conditional on total score

For the Rasch model the log likelihood ℓ_0 can be written as

$$\ell_0 = \sum_{i=1}^{n} x_i \log[P_i(\hat{\theta})/Q_i(\hat{\theta})] + \sum_{i=1}^{n} \log Q_i(\hat{\theta}) = \sum_{i=1}^{n} x_i d_i + C_r, \tag{4}$$

where x_i is the item response with $x_i=1$ for a correct response and $x_i=0$ for an incorrect response, $d_i=-\sigma_i$ where σ_i is the Rasch item parameter, and C_r is a constant given total score r; for fixed r all estimated abilities are equal. Molenaar and Hoijtink propose to evaluate response patterns conditional on total score r. The conditional probability of a response pattern in the Rasch model can be written as

$$P(\underset{\sim}{X} = \underset{\sim}{x} \mid \Sigma X_i = r) = \prod_{i=1}^{n} \varepsilon_i^{\,x_i} / \gamma_r(\underset{\sim}{\varepsilon}), \tag{5}$$

where $\varepsilon_i = \exp(d_i)$ and $\gamma_r(\underset{\sim}{\varepsilon})$ is the elementary symmetric function of order r, the sum of $\Pi\varepsilon_i^{x_i}$ over all patterns with $\Sigma x_i = r$.

In order to measure deviance of a response pattern, the probabilities of Equation 5 should be ordered from low to high which is equivalent to an ordering of $\Sigma x_i d_i$ in Equation 4 from low to high: Equation 5 gives the probabilities corresponding to Equation 4. On basis of this ordering the cumulative distribution for patterns can be computed. Next, one can determine whether a pattern falls below the 100α-percentile, where, for example, α is .05. Patterns below the 100α-percentile are denoted aberrant.

Complete enumeration of all response patterns is feasible when the number of items n is not too large. For large n the number of patterns $\binom{n}{r}$ becomes unwieldy for a large range of values r. Molenaar and Hoijtink therefore decided to find an approximate solution for the distribution of $M=\Sigma x_i d_i$ conditional on r. Using the first moments of M they approximate the distribution of M by a distribution based on the chi-square. They suggest to use this approximation except for $r=1$, $n-1$, and, possibly $r=2, n-2$. For the computations they need the ele- mentary symmetric functions $\gamma_r(\underset{\sim}{\varepsilon})$, $\gamma_{r-1}^{(i)}(\varepsilon)$, $\gamma_{r-2}^{(i,j)}(\varepsilon)$ and $\gamma_{r-3}^{(h,i,j)}(\varepsilon)$, where the notation $\gamma_r^{(i)}(\varepsilon)$ indicates the elementary symmetric function of order r on basis of item parameters $\underset{\sim}{\varepsilon}$, except the parameter with the index between parentheses. The chi-square approximation is adequate in most cases with 'reasonable' distributions of item parameters.

The computation of all elementary symmetric functions also becomes a formidable task with increasing n, and the accuracy of the computations diminishes. One of the alternatives, used by Kogut (1987), is simulation. Kogut simulated response patterns based on item parameter estimates $\hat{\underset{\sim}{\varepsilon}}$ in such a way that at least 200 patterns were obtained for each score level $r(r=1,\ldots,n-1)$. Conditional on r the patterns were ordered on basis of M and an approximate cumulative distribution was obtained.

Simulation studies like Kogut's might be very useful. It is quite possible that computation time can be kept within reasonable limits through sampling design optimization. One possibility is to search explicitly for a distribution of θ

which minimizes expected sample size under the constraint that there are at least n_r patterns for total score $r(r=1,...n-1)$.
Up to this point we have restricted ourselves to the computation of the fit statistic within the context of the Rasch model. The criticism of Molenaar and Hoijtink with respect to the standardization of ℓ_0 in Equation 1, is valid for other IRT-models as well. Their alternative - conditioning on total score - is attractive with other models as well because the question 'Is this response pattern extreme given total score' has relevance independently of the underlying response model. Instead of grouping examinees with respect to estimated ability one may group them according to total score (Yen, 1984). For tests of reasonable length this will also result in groups relatively homogeneous w.r.t. estimated ability. Actually, this idea was put forward by Van der Flier (1980) who formulated deviance of a pattern in terms of its probability of exceedence conditional on score level. However, he failed to distinguish between conditional and unconditional probabilities. He did not elaborate his ideas because he expected the computational problems to be unsurmountable. Molenaar and Hoijtink were the first to apply the conditional method.

In the Rasch model the probability of a response pattern given total score r does not depend on latent ability θ, and the item parameters $\underset{\sim}{\varepsilon}$ can be estimated conditional on total scores. It seems possible to apply the results of Molenaar and Hoijtink as an approximation in the context of other IRT-models when item parameters are adequately estimated and test length is large. In such a situation the probabilities correct might be approximated locally, i.e. given total score r, by Rasch item parameters. In the two-parameter model, for example, one can approximate the probability of a correct response to item i given estimated ability corresponding to total score r, $\hat{\theta}_r$, (Yen, 1984) by

$$\varepsilon'_{ir}/(1+\varepsilon'_{ir}) \simeq \Psi[a_i(\hat{\theta}_r-b_i)], \qquad i=1,...,n, \qquad (6)$$

where Ψ is the cumulative logistic, and a_i and b_i are the item parameters. Next, Rasch-parameter estimates ε'_{ir} can be rescaled so that their product equals one. One should remain aware of the fact that the approximation is not based on a conditional estimation method for item parameters. However, MML-estimation might also result in good item parameter estimates and with long tests even UML can provide adequate estimates. In the two-parameter model n-1 different Rasch models are possible. With the three-parameter model a complication arises as there may be scores below the pseudo-chance level.
Snijders (1988) suggested to simulate patterns with equal probabilities of occurrence for all patterns with a given value r. The sampled patterns must be

reweighted afterwards. With the method the lower tail of the target distribution might be approximated reasonably well.

The method requires the computation of $\gamma_r(\underset{\sim}{\varepsilon})$. Therefor an alternative was investigated. Patterns were generated for a fixed value r. Only the first occurrence of a pattern was computed: Sampling was done without replacement from the finite set of all possible patterns. The patterns were reweighted, and mean and standard deviation were computed. Computation of the elementary symmetric function was not needed because both the numerator and denominator in the evaluation of mean and variance contained this factor.

Several simulations with the proposed procedure were done in the context of the present study. Unfortunately, the results were not very promising. Due to the fact that the patterns are sampled with equal weights, the part of the distribution with the highest frequencies is underrepresented. It might be stated that sampling patterns results in an inaccurate estimate of the pattern distribution before reweighting, and the errors can be enlarged by the reweighting procedure. A possibly more accurate procedure is explored in the next section.

The distribution of *M* given *r*

Let us designate the target distribution $f_r(M)$; $f_r(M)$ is the distribution of M given r. This distribution can be rewritten as

$$f_r(M)=g_r(M)h_r(M) \tag{7}$$

with

$$g_r(M)= \binom{n}{r} \exp(M)/\gamma_r(\underset{\sim}{\epsilon})$$

and

$h_r(M)$: a discrete distribution which gives at M the proportion of vectors $\underset{\sim}{x}$ for which $\Sigma x_i d_i$ equals M; when each pattern corresponds to a different value M, $h_r(M)$ equals $1/\binom{n}{r}$.

So, $h_r(M)$ is the distribution in which all patterns have the same probability and allowance is made for the possibility that several response patterns have identical values M. The computational problems arises from the fact that $h_r(M)$ is a discrete distribution with a large number of values M. The computational task is alleviated when $h_r(M)$ can be approximated by a parametric distribution with a low number of parameters $\underset{\sim}{\alpha}$, $h_{r;\alpha}(M)$. In that case the target distribution can be approximated by

$$f_r(M)\propto \ \exp(M)h_r(M), \tag{8}$$

and its moments can be obtained by numerical integration.

Which families of distribution should be considered? The distribution family should be flexible. The approximation of the true distribution should satisfy even higher requirements than the approximation used by Molenaar and Hoijtink. For, this distribution is not the target distribution itself and accuracies in its approximation might show up enlarged in the approximation of the target distribution. Misspecification of the distribution's right tail can have large consequences for the accuracy of the approximation, due to the influence of the factor exp(M). Therefore, it was decided to take a distributional family with a finite range of values. The family of beta distributions $\beta(p,q)$ was chosen for the demonstration in this study. The distribution has a lower limit of zero and an upper limit of one, but this restriction can be circumvented easily using the four-parameter form

$$h_{r;p,q,\ell,h}(M) \propto (M-\ell)^{p-1}(h-M)^{q-1}. \tag{9}$$

It was decided to obtain the lower and upper limit directly from the original distribution of M:

$$\ell = \sum_{i=1}^{r} d_i - \delta \tag{10a}$$

$$h = \sum_{i=n-r+1}^{n} d_i + \delta \tag{10b}$$

where δ is a small constant, equal to half the step size used in the numerical integration. The first moment of the distribution is

$$\mu_r(M)=(r/n)\Sigma d_i. \tag{11}$$

Here we will always equal Σd_i to zero in order to constrain the latent Rasch scale; so $\mu_r(M)=0$. The variance, $\sigma^2_r(M)$, is also obtained easily using well-known results from sampling theory. Sampling without replacement from a finite population of size n gives

$$\sigma^2_r(M/r) = \frac{\sigma^2}{r} (d_i) \frac{n-r}{n-1} \tag{12}$$

and the wanted variance is obtained through multiplication of $\sigma^2_r(M/r)$ with r^2. With $\Sigma d_i=0$ scores r and n-r give identical means and variances. Further one should notice that, due to the restriction $\Sigma d_i=0$, results for score level r are equivalent with results for score level n-r in connection with a test with $d'_i=-d_i$.

Given the mean and variance, the remaining two parameters, p and q, can be obtained. Using Equation 8 one can obtain $f_r(M)$.

A computational example is given in Table 1. Due to the small number of items the example serves as a demonstration only.

Table 1. A computational example with two out of five items correct

I. distribution of d.

n=5, r=2, d=-1.0, -.4, 0.0., .2, 1.2

$\mu(d_i)=0.0$ $\quad \sigma^2(d_i)=.528$

II. $h_2(M)$

$\ell=-1.4$ $\quad h=1.4$

M	$h_2(M)$	pattern
-1.4	.1	11000
-1.0	.1	10100
-.8	.1	10010
-.4	.1	01100
-.2	.1	01010
.2	.2	10001+00110
.8	.1	01001
1.2	.1	00101
1.4	.1	00011

$\mu_2(M)=0.0$ $\quad \sigma_2^2(M)=r \times .528 \times (n-r)/(n-1)=.792$

III. approximation of $h_2(M)$ by a four-parameter beta-distribution (δ=0.0)

$\mu_2(M)=0.0 = (h-\ell)p/t+\ell$

$\sigma_2^2(M)=.792 = (h-\ell)^2 p(t-p)/\{t^2(t+1)\}$

$t=p+q$

$p/t=.5$ $\quad p(t-p)/\{t^2(t+1)\}=1.01$

$t= .25/.101-1.0=1.475$ $\quad p=q =.7375$

This is a symmetrical U-shaped beta-distribution.

IV. computation of $f_2(M)$

Compute $A = \int_{\ell}^{h} \exp(x)(x-\ell)^{p-1}(h-x)^{q-1}dx$ $\quad B= \int_{\ell}^{h} x\exp(x)(x-\ell)^{p-1}(h-x)^{q-1}dx$

$$C = \int_{\ell}^{h} x^2\exp(x)(x-\ell)^{p-1}(h-x)^{q-1}dx$$

approximated mean and variance of the target distribution $f_2(M)$:

$\bar{M} = B/A, \; s^2(M) = C/A-(B/A)^2$

In the present study it was decided to compute mean $\bar{M}$ and standard deviation $s(M)$ for the distribution $f_r(M)$ on the basis of Equation 8 and to compare the values $\bar{M}$ and $\bar{M}$-2$s(M)$ with the corresponding values obtained by an application of Kogut's sampling approach. The demonstration was based on a hypothetical forty-item test with the following *d*'s: -2.0, -1.5, -1.4, -1.3, -1.2, -1.0(.1)-.1, 0.0 (2×), .05(.05).7, .75 (2×), .8 (2×), .85 (2×), .9, .95, 1.0, where the notation -1.0(.1)-.1 designates the values -1.0, -.9, -.8,...,-.1. This distribution has a disadvantage in that several *d*'s and distances between *d*'s are equal (see Molenaar & Hoijtink, 1989). The results were, however, close to those for a set of more irregularly spaced, but in other respect similar *d*'s. In the simulation at least 200 patterns were generated for $1<r<n-1$. For $r=1$ and $r=n-1$ simulation outcomes were replaced by exact outcomes. The results are given in Figure 1.

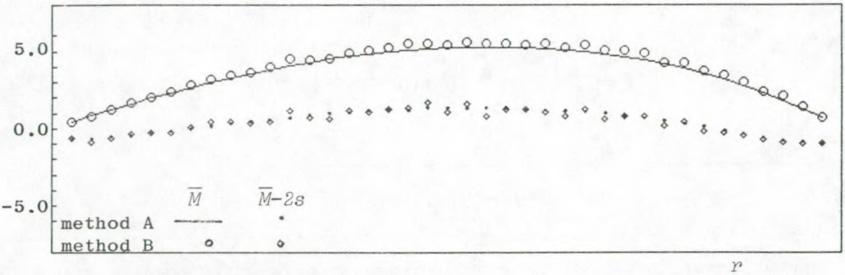

Figure 1. Estimates of means ($\bar{M}$) and lower limits ($\bar{M}$-2s) for the proposed method A, and the simulation method B.

The sampling results show small irregularities due to sampling error. It is clear that the sampling approach has its own drawbacks. The proposed approximation corresponds roughly with the sampling results. However, from the figure it is clear that the proposed method overestimates $\bar{M}$ for a large range of r. When accurate results are needed, the proposed method needs to be improved upon. One might think of using more moments in the determination of the parameters of distribution (9) or of using another distributional family like the generalized lambda distribution (Ramberg, Tadikamalla, Dudewicz and Mykytka, 1979). In a final test a variety of distributions should be used.

From the figure it is also clear that M varies less for more extreme values of r. One should already have suspected this from the relation between $\sigma^2_r(M)$ and r, derived from Equation 12. With relatively few or many correct responses there is less variation between values M and between pattern probabilities. In those cases one might wonder whether it is wise to order the patterns according to appropriateness - aside from possible problems with the accuracy of the ordering due to inaccurate item parameter estimates - and to label the least probable patterns deviant. The least probable patterns might differ only slightly in probability from the most probable patterns. Appropriateness measurement should be extended with a procedure by which to decide in which cases it is reasonable to order patterns according to deviance. One might borrow the mathematical approach from information theory. A relatively high redundancy (meaning that there is variation in probabilities) seems to be called for. Instead of redundancy one might use an index derived from $\sigma^2_r(M)$: if this variance is low, there is less variance in the conditional probabilities.

Discussion

Molenaar and Hoijtink (1989) have argued that the deviance of a response pattern should be considered conditional on the total score corresponding to the pattern. They restricted their arguments to the Rasch model. This author believes, with Van der Flier (1980), that conditioning on total score is sensible in other situations as well. In the present study a new approach to the estimation of the distribution of patterns was tried out. In this preliminary investigation only one set of item parameters was used. The results indicated that it might be worthwhile to try to improve the technique. The new approach might be useful as an alternative to simulations with large tests where it gives an approximation very fast. With small test exact results should be used, while at intermediate lengths the Molenaar-Hoijtink results are indicated.

Further, it was pointed out that blind computation of deviance might be misleading. The investigation should verify first whether patterns differ enough in probability. For small and large total scores the variation in pattern probabilities might be too small.

References

Drasgow, F., Levine, M.V., & Williams, E.A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology*, *38*, 67-86.

Drasgow, F., Levine, M.V., & McLaughlin, M.E. (1987). Detecting inappropriate test scores with optimal and practical appropriateness indices. *Applied Psychological Measurement*, *11*, 59-79.

Harnisch, D.L., & Linn, R.L. (1981). Analysis of item response patterns: questionable test data and dissimilar curriculum practices. *Journal of Educational Measurement*, *18*, 133-146.

Kogut, J. (1987). Detecting aberrant response patterns in the Rasch model. *Report 87-3*. Technical University Twente, Department of Education.

Levine, M.V., & Rubin, D.F. (1979). Measuring the appropriateness of multiple choice test scores. *Journal of Educational Statistics*, 4, 269-290.

Molenaar, I.W., & Hoijtink, H. (1989). The many null distributions of person fit statistics. *Psychometrika*, in press.

Rudner, L.M. (1983). Individual assessment accuracy. Journal of *Educational Measurement*, *20*, 207-219.

Sato, T. (1975). *The construction and interpretation of S-P tables*. Tokyo: Meiji Tosho (Japanese).

Snijders, T.A.B. (1988). Personal communication.

Van der Flier, H. (1977). Environmental factors and deviant response patterns. In Y.H. Poortinga (Ed.), *Basic problems in cross cultural psychology*. Amsterdam: Swets & Zeitlinger.

Van der Flier, H. (1980). *Vergelijkbaarheid van individuele testprestaties*. Lisse: Swets & Zeitlinger.

Wright, B.D., & Stone, M.H. (1979). *Best test design: Rasch measurement*. Chicago: Mesa Press.

Yen, W.M. (1984). Obtaining maximum likelihood trait estimates from number-correct scores for the three-parameter logistic model. *Journal of Educational Measurement*, *21*, 93-111.