KM 18(1985)
pag 107-114

TEST LENGTH AND ACCURACY OF PASS/FAIL DECISIONS IN CRM

Dato N. M. de Gruijter*

Summary

It has been demonstrated that Fhanér's (1974) approximate solution to the minimum test length problem, based on an indifference zone approach, is inaccurate. A minor modification discussed here will generally produce satisfactory results. The new procedure is extended to the situation with stratified random sampling, where an exact solution is not feasible.

1. INTRODUCTION

In criterion-referenced measurement mastery generally is defined in terms of the proportion π of items from the relevant item domain that the examinee can answer correctly. When this proportion equals or exceeds a standard π_0, the examinee is considered to be a master. Examinees with $\pi<\pi_0$ are non-masters. To determine whether an examinee has mastered the domain a random or stratified random sample of items from the item pool is administered. On the resulting test a minimum passing score, the cutoff score, is set. Only examinees with scores at least as high as this score are passed.

In pass/fail decisions two kinds of error can be made: passing true non-masters and failing true masters. The probabilities of these two kinds of error should be kept at a reasonably low level by choosing an appropriate number of items and an adequate cutoff score. This problem has been rephrased by Fhanér (1974) in terms of the problem to find the minimum test length for which certain requirement are met. He suggested a solution to the test length problem for tests with randomly selected items. This approximate solution has been criticized by Wilcox (1976) who demonstrated the feasibility of an exact solution.

Here it will be demonstrated that a combination of Fhanér's and Wilcox solutions is possible. In the case of the binomial model, the

* Bureau Onderzoek van Onderwijs, Boerhaavelaan 2, 2334 EN Leiden, tel. 071-148333, tst. 5392.

exact Wilcox procedure can still be used without problems. The new procedure can, however, also be used in connection with the generalized binomial model, in which an exact solution is not feasible. This will be demonstrated for a situation with stratified random sampling. First, Wilcox and Fhanér's solutions to the test length problem are reviewed.

2. THE SOLUTION OF WILCOX

In the Wilcox procedure an indifference zone around π_0 is specified from a lower bound π_1 to an upper bound π_2. The decision maker is indifferent with respect to the decisions concerning examinees with π-values in this zone. So π_1 represents the highest true non-master level and π_2 the lowest true master level. When random sampling of items is assumed, the probability of incorrectly passing a non-master at π_1 for a given number of items n and a given cutoff score c, $P_1(n,c)$, can be computed with the aid of the binomial model. In the same way the probability of incorrectly failing an examinee at π_2, $P_2(n,c)$, can be computed. In the Wilcox method as formulated by Hambleton and De Gruijter (1983) a maximum P^* is set to the probabilities of classification errors. The minimum test length is the smallest n for which the largest of the two error probabilities drops to a value equal to or lower than P^*, with a cutoff chosen so as to make the larger error probability as small as possible. In other words, the minimum test length is the smallest n for which

$$\min_c\{\max[P_1(n,c),P_2(n,c)]\}\leq P^*. \qquad (1)$$

The minimum test length can be obtained as follows. One takes a low trial value n. For this value $P_1(n,c)$ and $P_2(n,c)$ are computed for all possible values c. One verifies whether Inequality (1) is satisfied for one of the c-values. If not one of the c-values is satisfactory, the test length is increased by one. The procedure is repeated until the first n is obtained for which the inequality holds. Due to the fact that the binomial distribution is discrete, it is possible that Inequality (1) is not satisfied when the test length is again increased by one.

3. FHANER'S APPROXIMATION

In Fhanér's approximation the binomial distribution is approximated by a normal distribution. The probability of incorrectly passing a

π_1-level examinee on an n-item test with cutoff c is approximated as $\Phi(-z_1)$ with

$$z_1 = (c-n\pi_1)/n^{\frac{1}{2}}s_1, \tag{2}$$

where Φ denotes the normal distribution function and $s_1 = \pi_1^{\frac{1}{2}}(1-\pi_1)^{\frac{1}{2}}$. Similarly, the second error probability is approximated as $\Phi(-z_2)$ with

$$z_2 = (n\pi_2-c)/n^{\frac{1}{2}}s_2, \tag{3}$$

where s_2 equals $\pi_2^{\frac{1}{2}}(1-\pi_2)^{\frac{1}{2}}$.

By setting z_1 equal to z_2 one obtains the cutoff score

$$c = n(s_2\pi_1+s_1\pi_2)/(s_1+s_2), \tag{4}$$

which approximately minimizes $\max[P_1(n,c),P_2(n,c)]$ as a function of c, for fixed n.

The common value of z equals

$$z_m = n^{\frac{1}{2}}(\pi_2-\pi_1)/(s_1+s_2). \tag{5}$$

The maximum acceptable error probability P^* corresponds to a value of z, i.e. $P^* = \Phi(-z^*)$. Setting z_m in Equation (5) equal to this z-value and solving for n, the minimum test length is obtained with a c-value given by Equation 4. Generally the resulting value is not an integer, in which case n is raised to the first largest integer value. The factor $(\pi_2-\pi_1)/(s_1+s_2)$ is similar to Birnbaum's global information measure (Birnbaum, 1968), the difference being that Birnbaum assumed items of equal difficulty.

With respect to the approximate solution of Fhanér two problems arise. First, the accuracy of the normal approximation to the binomial migth be questioned. Secondly c in Equation (4) is a continuous function of π_1 and π_2 and consequently might take all possible values, in contrast with the real cutoff score. Fhanér noticed that the approximate solution tends to underestimate the minimum test length for a given criterion P^*. For that reason Wilcox rejected the approximation in favour of the exact solution.

4. A NEW APPROXIMATION

The error probabilities P_1 and P_2 for an optimal choice of c in the exact solution are highly irregular functions of n due to the discontinuous character of the raw score scale. The irregularity is nicely demonstrated in the numerical example of Hambleton and De Gruijter (1984), given in their second table. So, it is possible that the problems mentioned at the end of the previous section, are not so much due to the normal approximation as to the particular choice of the cutoff score. This possibility is explored in this section.

In this section the logistic approximation[1] is used, and c is restricted to integer values minus a continuity correction of 0.5. For example, when examinees with scores equal to 8 or higher are to be passed, c is set equal to 7.5. The procedure can be as follows. First, preliminary values c_0 and n are computed -- with the logistic instead of the cumulative normal distribution -- using Equations (4) and (5). Next, the two permitted values c on both sides of the preliminary value c_0 are considered; as an example, when the preliminary c_0 equals 17.2 the two permitted values are 16.5 and 17.5. For the two permitted values the probabilities $P_1(n,c)$ and $P_2(n,c)$ are approximated with the cumulative logistic distribution function, and compared using the minimax rule from the Wilcox procedure. When both c-values result in unacceptable error probabilities, n is increased by one. The procedure is repeated until a satisfactory combination of n and c is obtained.

For demonstrational purposes the new approximation has been used with π_1=.65 and π_2=.85. For n=8 and n=20 the exact error probabilities for these two π-values, given the correct c-values, are given by Hambleton and De Gruijter (1983). In Table 1 both the exact and the approximated probabilities are given for n=8 to n=20. The final c_0-values of the approximation are all equal to the c_0-values of the exact solution, when 0.5 is added to the c_0-values in the approximation. The old, normal approximation is not given; there z_1 (Equation 2) and z_2 (Equation 3) and, consequently, the two error probabilities P_1 and P_2 are set equal.

[1] The logistic curve $\{1+\exp(-Dz)^{-1}\}$ with scaling factor D=1.7 deviates nowhere more than 0.01 from the cumulative normal distribution $\Phi(z)$.

Table 1

Cutoff scores and probabilities of misclassification for $\pi_1=.65$ and $\pi_2=.85$

Test Length (n)	c both solutions	exact solution P_1	exact solution P_2	logistic approximation P_1	logistic approximation P_2
8	7	0.17	0.34	0.16	0.38
9	7	0.34	0.14	0.32	0.14
10	8	0.26	0.18	0.24	0.18
11	9	0.20	0.22	0.19	0.23
12	10	0.15	0.26	0.15	0.28
13	10	0.28	0.12	0.26	0.11
14	11	0.22	0.15	0.21	0.14
15	12	0.17	0.18	0.17	0.18
16	13	0.13	0.21	0.13	0.21
17	13	0.23	0.10	0.22	0.10
18	14	0.19	0.12	0.18	0.12
19	15	0.15	0.14	0.15	0.14
20	16	0.12	0.17	0.12	0.17

The largest discrepancy between the exact solution and the logistic approximation is less than 0.015 for $n \geqq 15$; most discrepancies are much smaller. When the critical error probability P^* is set equal to 0.17 both the exact and the approximate procedure result in a minimum test length equal to 19.

From the results in Table 1 and other results (for example, those from Tables 2 and 3, which will be discussed later) the new approximation seems fairly accurate. The approximation is a reasonable one, at least for cases where π_1 and π_2 are not too extreme or too far apart, and these cases are more likely to occur in criterion referenced measurement.

5. STRATIFIED RANDOM SAMPLING

The approximation seems more interesting in case of stratified random sampling. Stratified random sampling is frequently used instead of ordinary random sampling in order to diminish differences between alternative tests and to enhance accuracy of decision making.[2] The gain of stratified random sampling over random sampling increases with in-

[2] Of course, if item characteristics are known, tests consisting of randomly sampled items could be equated; in that case random sampling is not more than a possible item selection technique. In the present context the term 'random sampling' also entails that item characteristics are not known or not used.

creasing differences between stratum difficulties. It should also be clear that this advantage can only be fully exploited when the stratum difficulties are known. It will be assumed that this is the case.

Writing $\pi(i)$ for the average item difficulty of stratum i for an examinee with domain score π, and f_i for the relative size of the stratum, it is clear that

$$\pi \equiv \Sigma f_i \pi(i). \tag{6}$$

There are various possible stratified sampling strategies, one being proportional sampling. In proportional sampling the numbers of items selected from different strata are proportional to the relative sizes. The total score in proportional sampling is distributed according to the generalized binomial model with parameters $\pi(i)$. The expected total score equals $n\pi$ and its variance equals

$$\begin{aligned} \text{var}(X|\pi) &= n\Sigma f_i \pi(i)[1-\pi(i)] \\ &= n\pi(1-\pi) - n\text{var}[\pi(i)]. \end{aligned} \tag{7}$$

With stratified random sampling the two-term approximation to the generalized binomial model (cf. De Gruijter & Van der Kamp, 1984, Equation 6-41) could be used in order to obtain a cutoff score and error probabilities for a given choice of n. Alternatively, the simpler logistic approximation could be used, replacing the standard deviations for the binomial model in Equations 2-5 by the standard deviations for the generalized binomial model; in other words, s_1^2 should be set equal to $n^{-1}\text{var}(X|\pi_1)$, using Equation 7, and s_2^2 to $n^{-1}\text{var}(X|\pi_2)$.

As an example, take an item domain with three equally sized strata. As stratum difficulties for a π_2-level examinee the values $\pi_2(1)=.4$, $\pi_2(2)=.6$ and $\pi_2(3)=.8$ are chosen, i.e. π_2 is equal to .6. The difficulties for a π_1-level examinees are $\pi_1(1)=.22$, $\pi_1(2)=.4$ and $\pi_1(3)=.64$; in other words, π_1 is set equal to .42.

The division of the domain in strata can be neglected in favour of random item selection, which produces the results in Table 2. The approximation is fairly accurate, even better than the one in Table 1. Undoubtedly this is due to the fact that the π-values are less extreme in this example.

Table 2
Cutoff scores and probabilities of misclassification for π_1=.42 and π_2=.64 under simple random sampling

Test Length (n)	c both solutions	exact solution P_1	exact solution P_2	logistic approximation P_1	logistic approximation P_2
8	5	0.21	0.41	0.20	0.41
9	5	0.31	0.27	0.30	0.26
10	6	0.20	0.37	0.20	0.37
11	6	0.29	0.25	0.29	0.24
12	7	0.20	0.34	0.19	0.33
13	7	0.28	0.23	0.27	0.22
14	8	0.19	0.31	0.18	0.30
15	8	0.26	0.21	0.26	0.21
16	9	0.18	0.28	0.18	0.28
17	9	0.25	0.20	0.24	0.19
18	10	0.18	0.26	0.17	0.26
19	10	0.24	0.19	0.23	0.18
20	11	0.17	0.25	0.17	0.24

Table 3 gives the results for the stratified random sampling plan. Results are reported only for test lengths which are multiples of three, as these are the only possible test lengths in proportional sampling from the domain at hand. The approximation to the customary two-term approximation is fairly accurate.

Table 3
Cutoff scores and probabilities of misclassification for π_1=.42 and π_2=.64 under stratified random sampling

Test Length (n)	c both solutions	two-term approximation P_1	two-term approximation P_2	logistic approximation P_1	logistic approximation P_2
9	5	0.30	0.26	0.29	0.25
12	7	0.18	0.33	0.18	0.32
15	8	0.25	0.20	0.24	0.19
18	10	0.16	0.25	0.16	0.25

From a comparison of Tables 2 and 3 it is clear that stratified sampling only resulted in minor improvements in accuracy. For example, the two error probabilities for n=9 and c=5 are 0.30 and 0.26 in the two-term approximation to the generalized binomial model, and 0.31 and 0.27 for the binomial model. This result is not unexpected: only when the variation in item difficulties is very large, the gain from using the generilized binomial model is substantial.

6. DISCUSSION

It has been demonstrated that the normal approximation to the probabilities of misclassification in the indifference zone approach, which has been suggested by Fhanér, can be improved through a minor modification. As such, this result might not be very interesting while -- with some extra computational effort -- an exact solution can be obtained for the binomial model. The new approximation has, however, also been used with a stratified random selection plan, in which it is an alternative to the two-term approximation to the generalized binomial model. In principle it could be applied with nonrepresentative item selection plans, like the optimal item strategy suggested by Hambleton and De Gruijter (1983), as well. This could be achieved by substituting the relevant relative true scores for the domain scores in Equations 2-5.

REFERENCES

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In Lord, F.M., & Novick, M.R. *Statistical theories of mental test scores*. Reading, Mass.: Addison-Wesley.

De Gruijter, D.N.M. & Van der Kamp, L.J.Th. (1984). *Statistical models in psychological and educational testing*. Lisse: Swets & Zeitlinger.

Fhanér, S. (1974). Item sampling and decision-making in achievement testing. *British Journal of Mathematical and Statistical Psychology*, *27*, 172-175.

Hambleton, R.K., & De Gruijter, D.N.M. (1983). Application of item response models to criterion-referenced test item selection. *Journal of Educational Measurement*, *20*, 355-367.

Wilcox, R. (1976). A note on the length and passing score of a mastery test. *Journal of Educational Statistics*, *1*, 359-364.

Ontvangen: 1-2-85