Measurement and Research Department Reports 92-4

A Note on Quality Indices for Tests Used for Grading

T.J.J.M. Theunissen

Cito

Measurement and Research Department Reports 92-4

A Note on Quality Indices for Tests Used for Grading

T.J.J.M. Theunissen

Cito
Arnhem, 1992

Abstract

Two indices for test quality are presented. Both are based on so-called classification matrices whose m rows are obtained as conditonal score distributions over m partitionings of the latent continuum. The first one is the permanent of this matrix; the second one is based on the observation that occasionally such classification matrices can be expressed as convex combinations of permutation matrices.

Key words: permanent, convex sum of permutation matrices, conditional distribution.

Introduction

In classical test theory, the KR20 can be seen as an index for test quality. It's advantage is that a single number summarizes a lot of 'information'; it's disadvantage is population dependency. Similarly for generalizability theory and generalizability coefficients. In Item Response Theory, test quality is commonly expressed by the test information function with it's well-known advantages. Test information has no global summarizing feature, however. It might therefore be of interest to have, alongside test information, an index of quality that summarizes a lot of 'information' into a single number. The purpose of this note is to present two suggestions for such index, one of more practical and one of more theoretical interest. The indices can only be interpreted in conjunction with a grading system, however. Since defining a grading scale on the latent continuum is common practice in many situations, especially in education, this is not seen as a disadvantage. Despite the appearances of the matrix from which both are derived (to be shown below), neither index is a measure of association nor of agreement. In fact, neither index is a statistic so concern about sampling distributions is unnecessary. The rows of the matrix from which both are derived are based on conditional distributions. As will be shown, the numerical values of the indices are completely determined by the width of categories defined on the latent ability continuum and test information.

Assume as given an ability continuum, a partitioning of this continuum or of a part of it, a test consisting of IRT-calibrated items and known test information for this continuum or for the relevant part. The partitioning can, for example, take the form of an equal interval grading scale defined on the continuum or part of it. For each grade a conditional score distribution can be computed in the following manner. For the computation of the conditional distribution of the maximum likelihood estimator, given that $\theta = \theta^*$ (with I as test information at θ^*), use is made of the asymptotic property that it is $N(\theta^*, 1/I(\theta^*))$ distributed. Further, m points $\theta(i)$ are specified, each representing a latent class on the continuum. The border between classes i and i+1 is assumed exactly between $\theta(i)$ and $\theta(i+1)$. From this it follows that the proportion of students with true ability θ^* obtaining a maximum likelihood estimate on the test lower than the upper boundary of class i can be approximated by

$$\phi\{(\theta_i + \theta_{i+1})/2 - \theta^*\}I^{\frac{1}{2}},$$

where Φ denotes the cumulative normal distribution. It is furthermore assumed, that the test consists of items that are very well calibrated, so that I at any point of the ability continuum can be seen as a fixed quantity. Doing so for each of m grades and writing the result as m rows of a matrix results in what will be called here an m $\times$ m classification matrix C. The rows of C are conditional score distributions. Matrix C is stochastic and as will be shown further on, sometimes even doubly stochastic (both rows and columns sum to one).

First proposal for an index

There is a little known function of a matrix, with some resemblance to the determinant of a matrix, that, in conjunction with extreme forms of matrix C produces a number that can be interpreted as a quality index for a test, given a partitioning, for example, a grading system. This function is the permanent of a matrix, written as per(C) for square matrices (although a more general definition applies to any rectangular matrix and is written as Per(C)) and is defined as follows:

$$\mathrm{per}(C) = \sum_{\pi} c_{1\pi(1)}\, c_{2\pi(2)} \cdots c_{m\pi(m)}$$

This notation should be read as follows. Regard all π = m! permutation matrices of order m. For each permutation matrix take the corresponding elements in C and multiply these elements. Next, summate over all permutations (see Minc, 1978).

For two classification matrices C the value of the permanent is easily determined. The first case concerns the perfect test with error-free measurement; the second case concerns the worst possible test that does not discriminate at all (tests that allocate subjects systematically more frequently to improper grades than to the proper grades will be disregarded here). In an example with a grading system with three categories, these matrices look as follows:

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \text{ and } \begin{pmatrix} .333 & .333 & .333 \\ .333 & .333 & .333 \\ .333 & .333 & .333 \end{pmatrix}.$$

It is clear that per(C) for a perfect test will always be 1, since it is a function of the identity or first permuation matrix. So the (theoretical) maximum value of per(C) is 1 (per(C) raeches

a maximum for any permutation matrix; we are , however, only interested in classification matrices with dominant diagonal). A suggestion for a minimum value (although easily computed in the example above, it is obvious that the computational burden increases rapidly for larger matrices) can be derived from the van der Waerden conjecture for doubly stochastic matrices that states

$$per(C) \geq \frac{m!}{m^m}$$

with equality being obtained when all elements of C are equal to 1/m. (The conjecture was actually proven in the early eighties; for details, see Minc, 1988). Since the classification matrix for the worst possible test is not only stochastic but doubly stochastic, the van der Waerden result applies in this case.

Normally, the classification matrices between these two extremes will have a dominant diagonal, where the diagonal elements are preceded (followed) by monotonically increasing (decreasing) elements. If matrix C happens to be doubly stochastic, the minimum value of per(C) is therefore known and is completely determined by m, the number of categories on the ability continuum. However, in practice C will usually be only row-stochastic and the van der Waerden conjecture does not apply directly. A very large number of classification matrices, however, will have values for per(C) that are above the van der Waerden minimum. Each matrix C that has a smallest diagonal element greater than the m-root of Min per(C) has of necessity a permanent greater than the minimum, since the product of the diagonal elements of C is already greater than this minimum. For 3-square matrices this lowest diagonal element is 0.605, being the cubic root of 0.222222, the minimum of per(C) for 3-square doubly stochastic matrices; for 5-square matrices the minimum diagonal element should be 0.521 . These values are well within the reach of well-designed tests with sufficient test information in conjunction with the appropriate grade scale. It is conjectured, however, that the permanent of all row-stochastic matrices with dominant diagonal is greater than the van der Waerden minimum for doubly stochastic matrices of the same order (for 2-square matrices it can be proven; the proof is trivial). The conjecture is based on the following observation on the m! components of the permanent, each consisting of m fractions that are to be multiplied. First, regard the matrix filled with elements (1/m) and its permanent as a sum of m! products of m identical fractions (1/m). Next change two elements of the first row

by subtracting something from one off-diagonal element and adding this to the diagonal element. Those components in which the diagonal element appears will increase while those in which the changed off-diagonal element appears will decrease by exactly the same amount. The permanent remains the same. Next, repeat this procedure for the second row. Again some components of the permanent will change but now the effects of aggregating higher fractions in some components in conjuntion with aggregating lower fractions in some other components appears. Since within each component the fractions are multipled, the first effect is larger than the second effect and the permanent will incease. So for each stochastic matrix with at least two dominant diagonal cells it appears that the permanent is larger than the van der Waerden minimum for doubly stochastic matrices. No numerical counterexamples could be found.

The permanent as index for test quality has some interesting properties.

1. It is a summarizing index. Although low-order classification matrices can easily be compared by visual inspection, this is not the case for matrices of higher order.
2. Since the rows of the classification matrix are derived from conditional distributions, the index is population independent.
3. The ability to discriminate between tests increases with the quality of the tests. This can be easily seen if one realizes that the higher the value of the index, the more it will be determined by the product of the diagonal elements. The comparatively rapid increase of the product of higher fractions as compared to the product of lower fractions causes this effect.
4. The index is a measure of how succesful the test is in ranking the students. For example, given a test and two grading systems with different degrees of coarsenes in partitioning the latent ability continuum, the coarser system will produce the higher value for the permanent of the corresponding classification matrix. The difference in minimum value of the permanent reflects the fact that for coarser grading systems, it is easier to present a correct ranking of subjects than for less coarser systems. Given a fixed number of categories, it is also influenced by changes in the width of the categories. The index is grading system dependent. (It is also conjectured, that given a test and a fixed number of categories on a specified part of the ability continuum, the permanent in the case of equal intervals will be larger than that of all cases of unequal intervals, at least in the case of uniform information function).

Second proposal for an index

The case of equal interval partitioning and uniform test information opens another perspective on the ranking function of tests. Given the difficulty of obtaining exact uniform test information in practice, the following is, to that extent, of theoretical interest. With uniform test information, I in (1) is a constant. All class intervals are of equal size and the steps from one θ^* to the next are also of equal size and equal to the class intervals. As can be checked, this leads to symmetry in the columns with respect to the corresponding rows, making the classification matrix under these conditions a doubly stochastic matrix. Note that no knowledge about the actual distribution of ability nor about the actual value of test information is necessary. Since C is doubly stochastic, a result known as the Birkhoff theorem (Minc, 1988) or the Birkhoff-von Neumann theorem (Syslo, Deo, & Kowalik, 1983, p.339) applies in this case. The theorem states the following: if an m-square matrix C is doubly stochastic it can be expressed as

$$C = w_1P_1 + w_2P_2 + \ldots + w_qP_q$$

where the P's are m-square permutation matrices, the w's sum to 1, all w's being positive and $q \leq m!$.

To see this, define a bipartite graph G on matrix C with two subsets of nodes of equal cardinality m. The rows of C correspond to one subset of G and the columns of C correspond to the other subset of G. For every non-zero entry in C there is an edge in G; the weight of that edge is equal to the numerical value of the corresponding element in C. Next, find a perfect matching in G (a perfect matching is a set of pairwise disjoint edges covering all vertices). If C contains no empty cells, the maximum number of perfect matchings is m! (see Figure 1).

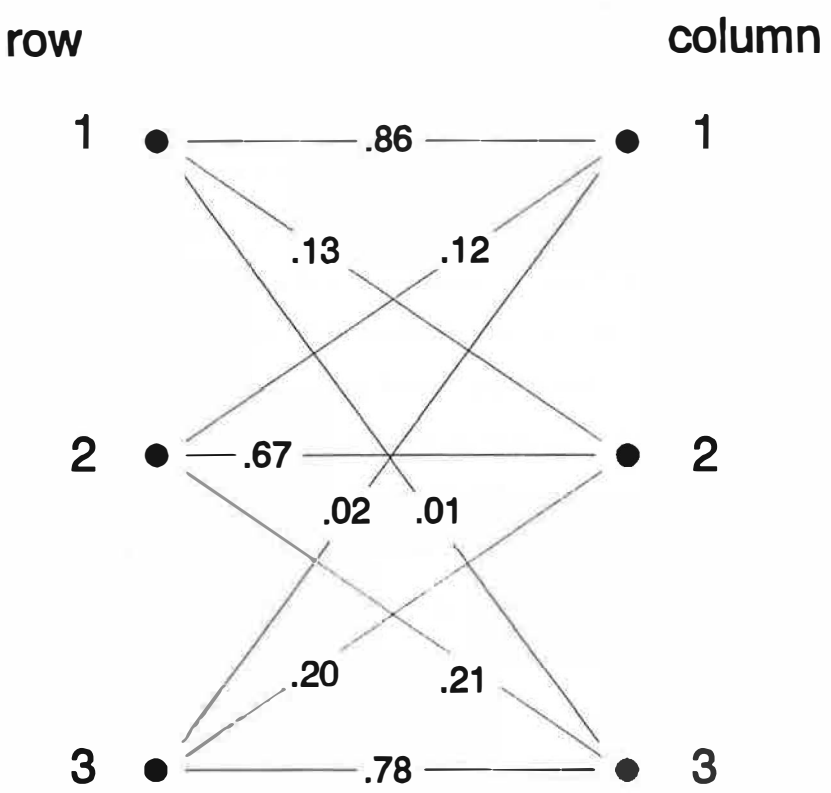

Figure 1. Perfect matching in a bipartite graph.

(Figure 1 is based on the following classification matrix, generated by the program Optimal Test Design; irregularities are due to the difficulty of obtaining exactly uniform test information given the bank used, and rounding in OTD):

and gets grade

$$\text{C} = \text{(deserves grade)} \begin{array}{c} \\ A \\ B \\ C \end{array} \begin{array}{c} \begin{array}{ccc} a & b & c \end{array} \\ \begin{pmatrix} .86 & .13 & .01 \\ .12 & .67 & .21 \\ .02 & .20 & .78 \end{pmatrix} \end{array}.$$

From Figure 1 it is easy to see that each perfect matching corresponds to a permutation matrix. It is also clear that given any set of five perfect matchings, the sixth one is determined, creating dependency. The weight of a perfect matching is the sum of the weights of its edges. Referring to figure 1, the weight of the perfect matching {(1,1), (2,2), (3,3)} is .86 + .67 + .78 = 2.31 . This perfect matching refers to the identity permutation matrix. Writing all m! permutation matrices in a certain order as, for example, below

$$\begin{pmatrix}1&0&0\\0&1&0\\0&0&1\end{pmatrix}\begin{pmatrix}1&0&0\\0&0&1\\0&1&0\end{pmatrix}\begin{pmatrix}0&1&0\\1&0&0\\0&0&1\end{pmatrix}\begin{pmatrix}0&1&0\\0&0&1\\1&0&0\end{pmatrix}\begin{pmatrix}0&0&1\\1&0&0\\0&1&0\end{pmatrix}\begin{pmatrix}0&0&1\\0&1&0\\1&0&0\end{pmatrix}$$

leads to equations as, starting with the first matrix,

$3w_1 + w_2 + w_3 + w_6 = 2.31$. Doing so for all matrices results in

$$\begin{pmatrix}3&1&1&0&0&1\\1&3&0&1&1&0\\1&0&3&1&1&0\\0&1&1&3&0&1\\0&1&1&0&3&1\\1&0&0&1&1&3\end{pmatrix}\begin{pmatrix}w_1\\w_2\\w_3\\w_4\\w_5\\w_6\end{pmatrix}\begin{pmatrix}2.31\\1.27\\1.03\\0.36\\0.33\\0.70\end{pmatrix}.$$

Two questions arise as regards (4). The first is: in how many ways can C be expressed in the form (4)? The second is: what is the least number $\beta(C)$ of permutation matrices whose convex combination equals C? According to Minc (1988), hardly anything is known about the first question. As regards the second question, several upper bounds have been presented (for details, see Minc, 1988). The best one known is given by

$$\beta(C) = h\left(\frac{m}{h} - 1\right)^2 + 1$$

where h is the index of imprimitivity. It will be shown that for classification matrices h will always be equal to 1. Because C contains probabilities, there will be no zero entries in C (although the actual values can become vanishingly small of course). This implies that the directed graph D associated with C is strongly connected, which happens to be the case when there is a path in D connecting any pair of vertices. Since there is no zero pattern in C, this is always the case. A path is a connection of two vertices in D and a cycle is a path connecting a vertex with itself. The greatest common denominator of the lenghts of all cycles in D is called the index of imprimitivity of D. C is called irreducible if its associated D is strongly connected and the index of imprimitivity of an irreducible matrix is equal to that of the associated matrix D (see Minc, 1988, Chapter 4). Since D contains loops (cycles of

length 1), h for D and therefore for C is equal to 1. All this means that $\beta(C)$ cannot exceed $(m - 1)^2 + 1$.

For the numerical example of C presented above this means an upper bound of $\beta(C) = 5$ or, in other words, C can be expressed as a combination of at most five permutation matrices. Choosing to set which w equal to 0 in the system of linear equations as presented above is guided by the following considerations. The choice should not affect diagonal elements nor, preferably, adjacent elements, since for regular classification matrices this is where most of the 'higher values' will be. The elements of the corresponding permutation matrix should have highest aggregate distance to the diagonal and the weight of the corresponding perfect matching (or sum of the corresponding fractions in C) should be lowest. These considerations lead to setting $w_5 = 0$ and solving for the rest with the following result:

$$.66\begin{pmatrix}1&0&0\\0&1&0\\0&0&1\end{pmatrix} + .20\begin{pmatrix}1&0&0\\0&0&1\\0&1&0\end{pmatrix} + .12\begin{pmatrix}0&1&0\\1&0&0\\0&0&1\end{pmatrix} + .01\begin{pmatrix}0&1&0\\0&0&1\\1&0&0\end{pmatrix} + .0\begin{pmatrix}0&0&1\\1&0&0\\0&1&0\end{pmatrix} + .01\begin{pmatrix}0&0&1\\0&1&0\\1&0&0\end{pmatrix} = \begin{pmatrix}.86&.36&.01\\.12&.67&.21\\.02&.20&.78\end{pmatrix}.$$

The interpretation is as follows: for this particular 3-grade system, the test behaves as a perfect test in 66 percent of the cases; in 32 percent (.20 + .12) it causes first order permutations, or switching adjacent grades and in only 2 percent it causes higher order permutations or switching several grades simultaneously. Since the range of any w is in principle from 0 to 1, it is tempting to regard w corresponding to the identity permutation matrix as an index of test quality. In that case the lowest value in the diagonal of the classification matrix is the upper bound of this index. It is interesting that both indices presented in this note are related to permutations and therefore to one of the purposes of a test: it should rank subjects correctly.

References

Minc, H., (1978). *Permanents*. In: Encyclopedia of mathematics and its applications, Vol. 6. Reading, MA: Addison-Wesley.

Minc, H., (1988). *Nonnegative matrices*. New York: Wiley.

Syslo, M.M., Deo, N., & Kowalik, J.S., (1983). *Discrete optimization algorithms*. Englewood Cliffs, NJ: Prentice-Hall.

Recent Measurement and Research Department Reports:

91-1 N.D. Verhelst & N.H. Veldhuijzen. A New Algorithm for Computing Elementary Symmetric Functions and Their First and Second Derivatives.

91-2 C.A.W. Glas. Testing Rasch Models for Polytomous Items: With an Example Concerning Detection of Item Bias.

91-3 C.A.W. Glas & N.D. Verhelst. Using the Rasch Model for Dichotomous Data for Analyzing Polytomous Responses.

91-4 N.D. Verhelst & C.A.W. Glas. A Dynamic Generalization of the Rasch Model.

91-5 N.D. Verhelst & H.H.F.M. Verstralen. The Partial Credit Model with Non-Sequential Solution Strategies.

91-6 H.H.F.M. Verstralen & N.D. Verhelst. The Sample Strategy of a Test Information Function in Computerized Test Design.

91-7 H.H.F.M. Verstralen & N.D. Verhelst. Decision Accuracy in IRT Models.

91-8 P.F. Sanders & T.J.H.M. Eggen. The Optimum Allocation of Measurements in Balanced Random Effects Models.

91-9 P.F. Sanders. Alternative Solutions for Optimization Problems in Generalizability Theory.

91-10 N.D. Verhelst, H.H.F.M. Verstralen & T.J.H.M. Eggen. Finding Starting Values for the Item Parameters and Suitable Discrimination Indices in the One-Parameter Logistic Model.

92-1 N.D. Verhelst, H.H.F.M. Verstralen & M.G.H. Jansen. A Logistic Model for Time Limit Tests.

92-2 F.H. Kamphuis. Estimation and Prediction of Individual Ability in Longitudinal Studies.

92-3 T.J.H.M. Eggen & N.D. Verhelst. Item Calibration in Incomplete Testing Designs.