Review of Psychology,
2009, Vol. 16, No. 2, 77-85 UDC 159.9

Scoring of multiple choice items by means of internal linear weighting

GREGOR SOČAN

Four scoring methods for multiple choice items were compared: number-right scoring, guessing correction, homogeneity analysis and weighting of choices according to their correlation with the number right score. The four methods were compared according to the validity, reliability and dimensionality of the resulting scores. Comparisons were based on simulated data, where the proportion of responses due to guessing and the discrimination power of options were controlled. The results showed an inferior performance of the guessing correction and a good performance of the proposed correlation-weighting, especially when traditional assumptions about multiple choice items were violated.

Key words: multiple choice, scoring, weighting, simulation

Multiple choice (MC) is a popular and sometimes virtually inevitable item format in the area of knowledge, aptitude and ability testing, or, more generally, in the testing areas where the maximal achievement is of interest and the response speed is not of cardinal importance. For simplicity, we shall use the term "ability" as a designation for the measured trait. A typical MC format consists of *m* offered alternatives, one of which is correct or at least the best, while the remaining alternatives ("distracters") are false. This includes the true-false format as a special case, *m* being equal to two in this case. We shall not treat explicitly MC formats with more than one correct alternative per item, although the principles discussed in the paper can be generalized to such item types as well.

Multiple choice tests have two specific features. First, the nature of a response to a test consisting of MC items is primarily categorical: an examinee responds by returning a pattern of non-numerical statements. The problem of transforming this pattern into a meaningful numerical value (usually termed as the test score), expected to be related to the level of the person's ability, is therefore not trivial. Of course, the simplest method of scoring is simply to count the number of correct choices made (we shall use the term "number-right score" for this number in the sequel). Obviously, this scoring method uses just a part of the available information on the actual response: if the examinee did not choose the correct option, it makes no difference whether the item was either omitted or attempted; and if it was attempted, it makes no difference which of the incorrect options was chosen. The second peculiarity of the MC items is that examinees who do not know the correct answer may still attempt to score a point by means of guessing. For a thorough discussion of the psychometric effects of guessing the reader is referred to, for instance, Lord and Novick (1968) or Zimmerman and Williams (2003). Here it is enough to stress that guessing generates an undesired part of the test score variance.

The aim of this paper is to consider some techniques that attempt to improve the estimation of examinee's ability by means of weighting the choices. We shall treat the problem from a perspective of a practitioner analyzing samples of moderate size at best, possibly using an *ad hoc* test where the main interest is on the particular group of examinees. University or high-school teachers performing regular examinations may serve as a typical example. For that reason we also deal only with the simplest administration technique, namely, the examinee marks one of the alternatives, which is the only correct alternative. Alternative administration techniques include, among others, elimination of one or more wrong options (Coombs, 1953), "answer-until-correct" (Poizner, Nicewander, & Gettys, 1978) and subset selection (Gibbons, Olkin, & Sobel, 1997). Apart from the widespread use of marking the single correct alternative, use

Gregor Sočan, Department of Psychology, University of Ljubljana, Aškerčeva c. 2, SI-1000 Ljubljana, Slovenia.
E-mail: gregor.socan@ff.uni-lj.si (the address for correspondence).

Acknowledgements:

The author wishes to thank the anonymous reviewer and the editor for their useful comments on the first version of the paper.

of this administration technique may be expected to keep the proportion of examinees responding contrary to instructions relatively low. As Cross and Frary (1977) found, in practice examinees often do not answer according to instructions whenever these get slightly complicated. This problem may be expected to become worse in a stressful situation like university exams.

A review of the literature shows that the problem of item scoring (including MC item scoring) in the context of the classical test theory received more attention in the past compared to more recent times. In their classic monographs on psychometrics, both Gulliksen (1950) and Lord and Novick (1968) devoted a whole chapter or a part of a chapter, respectively, to the problem of scoring, with special emphasis on the guessing correction. More recently, Nunnally and Bernstein (1994) mostly limit their discussion of scoring to the problem of guessing. McDonald (1999) makes just a short comment on MC scoring at the end of his textbook, in a chapter devoted to advanced topics. In the comprehensive Educational measurement (Brennan, 2006), the issue of scoring algorithms applicable to MC items is practically absent, although there is some discussion on some other aspects (for instance, computerized scoring). Finally, de Gruijter and van der Kamp (2008) do not mention the scoring at all and take the simple sum score (or number-right score, respectively) for granted.

One reason for the apparently diminishing interest for the scoring issues may be the shift towards the item response theory that has emerged in the psychometric theory. In recent decades several item response models were proposed which present an elegant way of the optimal use of the information contained in responses to MC items, for instance Bock (1972, 1997), Thissen and Steinberg (1984, 1997), Revuelta (2005), see also Thissen and Steinberg (1986). Unfortunately, because of a large number of parameters, these models do not always seem to be appropriate. For instance, in the Thissen and Steinberg model, the number of independently estimated parameters is 3m-1, where m is the number of response options, which makes the estimates unstable unless the sample is very large.

A simple a priori weighting system: correction for guessing

In situations where the IRT models are not considered appropriate, either because of sample size considerations or for other reasons, we need to resort to classical scoring, that is, computing the test score as a linear combination of appropriately recoded item responses. The simplest differentially weighted score is the guessing-corrected score, also known as formula score in some sources (for instance, Lord & Novick, 1968). In addition to the number of correct choices, the guessing-corrected score also takes into account whether the examinee chose one of the wrong options or omitted the item. In the usual application of the correction for guessing, a correct response is scored with 1 point, an omitted item with 0 points, and an incorrect response with -1/(m-1) points, where m equals the number of alternatives, including the correct one. It is obvious that the expected score for an examinee whose answers are based purely on guessing is equal to zero, and the score of an examinee who has not chosen any incorrect alternatives is equal to the number of correct responses. This scoring technique is aimed to produce test scores equaling to the number of correct answers that a particular examinee actually knows and to decrease the part of the test score variance due to guessing. The validity of the correction for guessing depends on two assumptions: first, correct response options are equally attractive as the incorrect ones, and second, all incorrect responses are due to guessing. The latter assumption states that no incorrect options were chosen because of some systematic misinformation, incorrect scoring key or similar external factors. The presence of partial information (when an examinee is able to eliminate some of the incorrect alternatives) does not invalidate the use of the guessing correction. Finally, it should be noted that other formulas for guessing correction have been proposed (see Gulliksen, 1950, or Reid, 1977), but the procedure described above seems to prevail both in practice and in the literature.

The opinions about the use of guessing-corrected scores differ widely. Lord and Novick (1968) do not recommend it for calculation of item difficulty, but on the other hand they have a more positive stance with regard to its use in test score computation. They stress that the "formula score" is an unbiased estimator of the number of items examinee actually knows and that small gains in validity can be expected when there are omitted responses.

Lord (1975) argued that the guessing-corrected score should be more reliable and valid than the number-right score. Further, if some answers are omitted, the guessing-corrected score "is always a better estimator of the examinee's standing on the trait measured" (p. 9). However, the reliability advantage depends on the validity of the assumption that examinees adapt their testing behavior to maximize their expected score depending on the scoring procedure. A subsequent empirical study by Cross and Frary (1977), however, showed that in real testing situations even university students may fail to choose an appropriate answering strategy and to answer exactly according to instructions. The differences in validity and reliability coefficients between number-right scores and guessing-corrected scores were negligible in the Cross and Frary study.

Burton (2004, 2005) argued in favor of the guessing correction, noting that it may improve the ratio of the standard error of measurement to the expected score range, therefore improving the measurement accuracy of the test. He did not, however, provide any empirical supportive evidence.

On the other hand, Nunnally and Bernstein (1994) seem to have a more critical opinion. They suggest that, instead of using the correction, examinees are instructed to answer

to all items, regardless of their confidence in the selected option. They also believe that the guessing-corrected score probably overestimates the number of correct answers that a particular person actually knows.

In a simulation study, Frary (1982) compared six modes of scoring MC items, including the number-right score and the guessing-corrected score. The remaining four modes included a modified item administration procedure. The latter modes turned out to be slightly more reliable and valid compared to the number-right score, however, in Frary's opinion the differences were not large enough to recommend replacing the number-right score with an alternative.

In a recent empirical study, Alnabhan (2002) compared three scoring methods in the context of a university examination, while controlling for guessing-proneness. The guessing-corrected scores had a larger coefficient alpha than the number-correct scores, but the difference was notable only in the group with the higher risk-taking level. On the other hand, the guessing-corrected scores had a lower predictive validity than the number-correct scores in the same group. In the group of students with a low risk-taking level the values of both coefficients was just slightly higher (.03 and .04, respectively) for the guessing-corrected scores.

The empirical evidence about the usability of the correction for guessing seems inconclusive. Lord's (1975) warning against putting too much value to comparisons of reliability measures like coefficient alpha should also be remembered. Namely, since alpha is just a lower bound to the actual reliability, a higher alpha does not necessarily imply higher reliability. Further, in Lord's opinion a reliability comparison only makes sense if the measured trait remains the same, but the use of the correction for guessing might change the nature of the trait measured.

Empirical weighting

Attempts to weight alternatives empirically go back at least to Thurstone (1919), who proposed assigning regression weights to both the sum of correct and the sum of incorrect answers by using some external criterion. However, later psychometricians were less enthusiastic about empirical weighting on the item level. Lord and Novick (1968) mentioned the possibility of a differential weighting of incorrect responses, but conclude that it may not be worth the effort. Nunnally and Bernstein (1994) warned quite emphatically against an uncritical use of optimized weighting of any kind and advocated equal weighting, mostly for the reason of stability. The paper of Wainer (1976) may have been influential in shaping such an "anti-weighting" opinion. Wainer showed that replacing an optimally weighted linear combination with an unweighted sum results in a negligible loss of predictive accuracy if the optimal weights are relatively close to each other. Specifically, if k optimal regression weights are randomly distributed on the interval [.25 .75], then the loss of the explained variance, resulting from replacing the true weights with .5, will be less than $k/96$. This "equal weights theorem" is often stated as an argument against using weights in calculating summary variables like total test scores. The prevalence of this view is reflected in the fact that the issue of weighting item responses has been practically missing from recent basic psychometric literature. One of the few papers that explicitly argued in favor of using optimal weights was the one by Hofstee, Ten Berge and Hendriks (1997), who recommended using principal components instead of simple sum scores in personality testing. These authors, however, only treated the personality questionnaires but not maximum performance tests.

Empirical evidence seems to support critical views. For example, using a large sample of students answering a vocabulary and a mathematical test, Kansup and Hakstian (1975) compared several scoring methods, including a priori »logical« weights and confidence weights. They found that although the logical weighting, contrasted to number-correct scoring, increased coefficients alpha, it failed to increase the test-retest reliability and the criterion validity. In fact, the weighted responses were in some cases notably less stable and valid. The authors concluded that the results were »disappointing to those who believe that substantial additional information can be learned about subjects' performance potential by going beyond conventional 0-1 scoring«. (p. 228).

Still, it may be premature to discard empirical weighting altogether, since there are several different ways to get the weights. Weights can be based on the following criteria:

1. subjective judgment of the test constructor on the degree of "falseness" of each alternative,
2. correlation with an external criterion, or
3. relation to an internal criterion.

The first two criteria may be simpler from the computational viewpoint, but they are difficult to be applied in practice. We only deal with the third criterion here.

The technique that seems to be the optimal alternative to categorical IRT models in the framework of the classical test theory is homogeneity analysis, also known as optimal scaling or multiple correspondence analysis. Homogeneity analysis is a multivariate method that can be seen as the variation of principal component analysis for categorical variables (Gifi, 1990). It transforms a set of categorical variables into one or more uncorrelated numerical variables which optimally summarize the information contained in the original categorical variables. As Greenacre (2007) defines its criterion, it assigns values to each category of the analyzed variables so that the average squared correlation is maximized between the scaled observations for each variable and their sum. This is equivalent to say that a minimum is sought for the variance between item scores within each respondent, averaged over sample – therefore, a maximally homogeneous set of quantified categorical variables is sought. A detailed treatment of homogeneity analysis is available in,

for instance, Gifi (1990), Michailidis and de Leeuw (1998) and Greenacre (2007). Gifi (1990) also presents a detailed example of a homogeneity analysis application for scaling multiple-choice items.

Surprisingly enough, homogeneity analysis as an item scoring technique has not found its way to either psychometric textbooks or practice. One of the few textbook authors proposing homogeneity analysis as a general approach to scoring multicategory items is McDonald (1999), although he mentions it just briefly in the context of advanced psychometric topics. We do not know of any empirical evaluation of homogeneity analysis as a scoring technique for MC items, but it can be expected to entail two problems that mirror the problems of the principal component analysis used in item analysis of numerically scored items:

1. the scaling weights may be unstable because of the sampling error;
2. we have no guarantee that the first dimension will be equivalent to the trait we wish to measure.

It thus seems that a more robust approximation to homogeneity analysis would be desirable. It is well-known that the uncorrected item-total correlation is a first-order approximation to the first principal component (see, for instance, Hofstee, Ten Berge & Hendriks, 1997, p. 901). Analogously, we propose that the test score be computed as the sum of the values of dummy variables, corresponding to each category, each of them weighted with the correlation coefficient between the dummy variable and the total number-right score. To our knowledge this weighting system has not been evaluated yet. Compared to the homogeneity analysis weighting, the proposed "correlation weights", as we shall call them in the sequel, are not optimal in the sense described above; however, we can be sure that the resulting test score will closely correspond to the measured ability.

The problem of the study

We shall empirically compare four approaches to scoring MC items within the framework of the classical test theory, that is, based on computing linear combinations:

1. the number right score (NR),
2. the guessing-corrected score (GC),
3. the first dimension obtained by homogeneity analysis (HA) and
4. the sum of the dummy variables weighted by the correlation weights (CW).

We shall compare the four techniques according to the following criteria:

1. validity, defined as the correlation with the latent ability underlying actual item responses,
2. the greatest lower bound to the reliability (GLBR),
3. the degree of unidimensionality, defined as the proportion of the common variance, explained by the first common factor (we shall use the term "explained common variance (ECV)" in the sequel).

METHOD

Data generation

Data were obtained by means of simulation. The following experimental conditions were involved:

1. presence of a distracter (an alternative scored as incorrect) with a positive discrimination vs. all alternatives discriminating negatively, except the correct one;
2. degree of guessing: high (5% questions omitted) vs. moderate (25% omitted); the average proportion of correct answers was 50% in both cases.

Therefore, four combined experimental conditions were used. For each of the four combinations, 1000 data matrices were generated as follows. First, $N = 100$ ability values were sampled from the normal distribution. The random number generator incorporated in MATLAB 5 (1998) was used for this purpose. Then the probability of choices of the response options were determined for each person × item combination according to Bock's (1972, 1997) nominal categories model. This model is a classical item response model for the analysis of categorical items. It models the item response as a function of two parameters, the first one corresponding to the frequency of the choice of a particular option, and the second one being related to the discrimination power and the ordering of the response options. The values of item parameters were chosen so that the desired expected proportions of the correct and the omitted answers were obtained. An item response model was used for data generation because of its relative simplicity and to make the data generation procedure as neutral as possible with regard to the scoring algorithms. Finally, actual responses of the examinees to the items were determined. This was carried out by means of the random number generator, so that the actual probability of choosing an option was equal to the probability determined by the Bock model.

The length of the test was 20 items with 4 alternatives; sample size was 100. The parameter values for the nominal categories model are given in the Appendix.

Algorithms

Homogeneity analysis algorithms as implemented in statistical packages like SPSS are often based on an iterative procedure called alternating least squares (for details see Gifi, 1990). However, in our case, where only the first dimension was of interest, the explicit solution algorithm (see, for instance, ten Berge, 1993, pp. 66-67) was considered to be more convenient from the computational viewpoint. This means that the HA weights were computed by means of a

closed-form solution rather than using an iterative estimation procedure. This choice did not affect the values of the obtained weights.

The greatest lower bound to reliability (GLBR), which is the best possible lower-bound estimate of the sample reliability, was computed using a minimum trace factor analysis algorithm proposed by ten Berge, Snijders and Zegers (1981).

Explained common variance (ECV) was determined with the minimum rank factor analysis (ten Berge & Kiers, 1991). The one-factor solution was computed and then the measure of unidimensionality ECV was computed as the 100 × variance explained by the first common factor divided by the total common variance (for a discussion on using ECV as a unidimensionality measure see, for instance, ten Berge & Sočan, 2004). We have chosen this particular measure because the proportion of common variance, which is explained by the first common factor, is in our opinion the most straightforward indicator of the unidimensionality of a group of variables. The minimum rank factor analysis was used because it is the only factor analysis method which makes possible to compute the proportion of the explained *common* variance.

All simulations and computations were performed using MATLAB 5 (1998).

RESULTS

Validity

The correlation coefficient between each of the four scores and the value of the latent ability was taken as the measure of validity. Analysis of variance with repeated measures on the scoring method factor showed statistically significant effects of scoring method ($F(3,11988) = 6005.9$; $p < .001$, partial $\eta^2 = .60$), scoring × condition interaction ($F(9,11988) = 2679.4$; $p < .001$, partial $\eta^2 = .67$) and condition ($F(3,3996) = 1010.5$; $p < .001$, partial $\eta^2 = .43$). Of course, because a high power is easily reached in simulation studies like this one, the low p values should not be surprising. Figure 1 describes the distribution of validity coefficients for each scoring method (NR, GC, CW and HA, respectively) and for each combination of conditions. First, we can note that validity of each score type was at least a bit lower when the guessing level was high (conditions II and IV) than when it was moderate (conditions I and III). Second, if none of the incorrect alternatives discriminated positively (conditions III and IV), the validity of NR and GC scores was higher than in the situation with one positively discriminating alternative besides the scored-as-correct alternative. The presence of such a distracter made both the CW scores and the HA scores notably more valid than the NR and the GC scores. On the other hand, in conditions with no positively discriminating distracters the CW scores were about as valid as the NR scores, and the HA scores were less valid – in the condition with a high guessing level they were even less valid than the GC scores. In all conditions, median HA validity coefficients were lower than median CW coefficients, and median GC validity coefficients were lower than median NR coefficients.

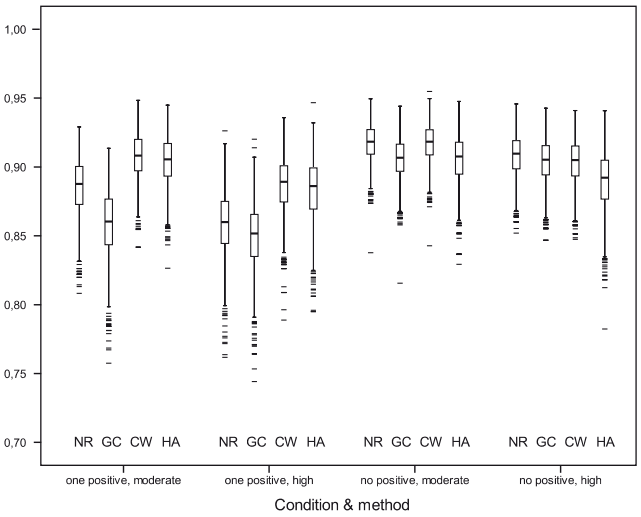

Figure 1.
Box-and-whiskers plots for the validity coefficients

Note. "one positive" = one distracter with positive discrimination, "no positive" = no distracters with positive discrimination, "moderate"/"high" = guessing levels. NR = number-right, GC = guessing correction, CW = correlation weights, HA = homogeneity analysis.

Following Wainer's (1976) reasoning, we computed the increment or loss, respectively, of the explained variance related to the latent ability – score relationship, for cases when a method other than the number-right scoring was used. We computed the difference in explained variance for

Table 1
Descriptive statistics for explained variance increment/loss compared to NR scoring

Condition	Score	Mean	Median	SD
	GC	-.048	-.047	.011
I	CW	.039	.039	.017
	HA	.033	.032	.025
	GC	-.015	-.015	.005
II	CW	.049	.049	.019
	HA	.044	.044	.034
	GC	-.021	-.021	.008
III	CW	-.001	.000	.009
	HA	-.022	-.020	.017
	GC	-.008	-.007	.004
IV	CW	-.008	-.008	.008
	HA	-.033	-.031	.024

each simulated sample and then computed descriptive statistics. Table 1 presents the results. A positive mean/median value indicates that a particular score type is more valid than the number-right score and vice versa. We can see that, depending on the condition, we can expect to lose between 0.8% and 4.8% explained variance if we use the guessing-corrected score instead of the number right score. Further, in conditions with positively discriminating distracters we can gain more than 4% of explained variance by using either CW or HA scores instead of the number-right scores. But in conditions with no positively discriminating distracters the HA scores result in an expected loss of more than 3%. The loss values of HA scores had the highest standard deviation, which means that the effectiveness of the HA scoring relative to the NR scoring was the least predictable among the three compared methods.

Reliability

As before, a repeated measures ANOVA indicated statistically significant effects of the scoring method ($F(3,11988) = 52097.9$; $p < .001$, partial $\eta^2 = .93$), scoring × condition interaction ($F(9,11988) = 5720.8$; $p < .001$, partial $\eta^2 = .81$) and condition ($F(3,3996) = 909.9$; $p < .001$, partial $\eta^2 = .41$). Figure 2 shows that the HA scores, followed by the CW scores, had the highest median values of GLBR, while the GC scores had the lowest median GLBR's in all four conditions. As in case of validity, the advantage of the CW and the HA scores was higher when a positively discriminating distracter was present. For all four score types the median reliability estimate was lower when the guessing level was higher.

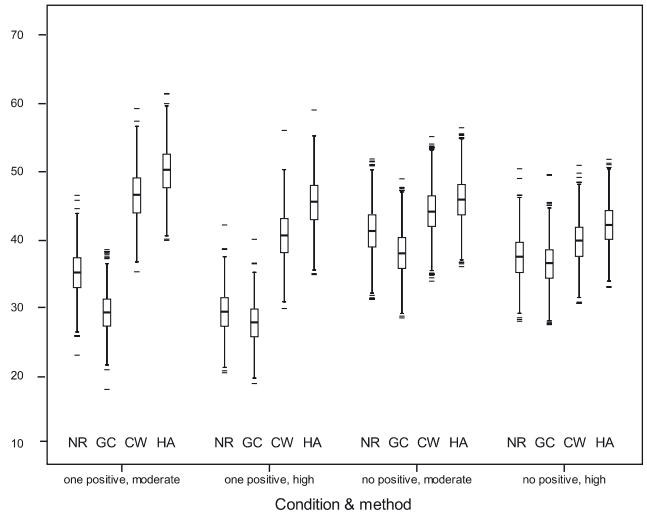

Figure 2.
Box-and-whiskers plots for the reliability coefficients
Note. See explanations next to figure 1.

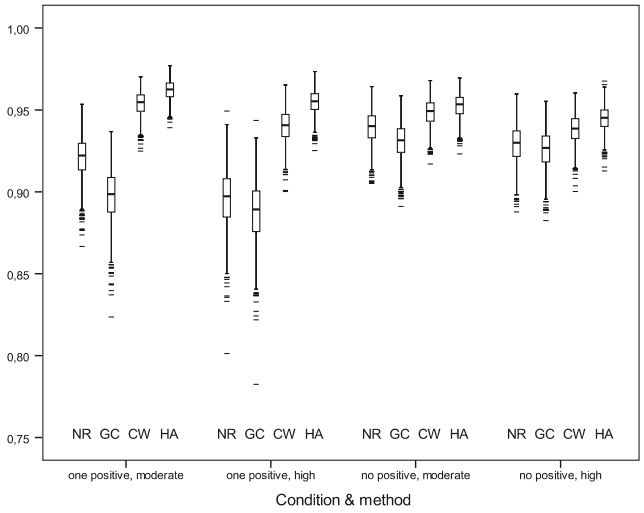

Figure 3.
Box-and-whiskers plots for the explained common variance percent
Note. See explanations next to figure 1.

Unidimensionality

Again, the ANOVA results show statistically significant effects of the scoring method ($F(3,11988) = 97202.2$, $p < .001$, partial $\eta^2 = .96$), scoring × condition interaction ($F(9,11988) = 9262.6$, $p < .001$, partial $\eta^2 = .87$) and condition ($F(3,3996) = 729.3$, $p < .001$, partial $\eta^2 = .35$). Figure 3 shows a situation very similar to the previous one: when the items were scored by HA, the median proportion of the common variance explained by the first common factor was the highest. The rank order of the scoring methods was the same in all four conditions: HA produced the most unidimensional scored items and the GC scoring the least unidimensional items. A higher amount of guessing decreased unidimensionality as well as validity and reliability. The presence of a positively discriminating distracter increased ECV for the CW scores and the HA scores, while it decreased the ECV for the remaining two score types.

DISCUSSION

The simulation technology makes possible a direct estimation of construct validity in sense of the correlation with the latent trait underlying the test responses. In our study, the most important results are those related to validity, which also present the added value of our study compared to most previous studies.

In general it seems safe to conclude that the use of guessing correction can not be recommended. The guessing-corrected scores were less valid, less reliable and less unidimensional than the simple number-right scores in all

conditions, even when the guessing level was high (that is, when only 10% of those who did not know the answer omitted the item). In some cases, especially when the guessing level was moderate (that is, when one half of those who did not know the correct answer omitted the item), these scores were notably worse compared to the number-right scores. The bad performance of the guessing-corrected scores can be easily understood from the weighting perspective. We argue that the guessing correction is quite a peculiar weighting scheme. In statistical weighting systems (regression weights and principal component weights being typical examples) the weights reflect the strength of relationship with some criterion, possibly corrected for interrelations with other variables. The guessing correction weights, on the other hand, have no relation to the degree of "falseness" of an alternative. Further, an omission receives a higher weight (namely zero) than an incorrect choice, although the factors leading to either omission or an incorrect choice are probably related to personality characteristics, instructions and other irrelevant factors. It follows that the guessing correction involves an implicit assumption, which seems to have been overlooked, namely that all incorrect alternatives are equally incorrect and that no items were omitted. However, if all examinees respond to all items, the guessing correction is not needed any longer anyway.

The performance of the homogeneity analysis was not encouraging. The advantage of the HA scores is their superior internal consistency: these scores had the highest average values of both reliability measures and unidimensionality measures. This is not surprising. High unidimensionality can be safely related to the very criterion of homogeneity analysis, namely, to get the weights producing the most homogeneous set of quantified variables. Further, these weights also maximize the value of the coefficient alpha for the sum score, analogously to the principal component weights (for details see, for instance, Greenacre, 2007). Unfortunately, the average validity coefficients obtained by the homogeneity analysis scores were lower than the coefficients for the correlation-weighted scores, and in conditions with no positively discriminating distracter they were even lower than the coefficients of the number-right scores. It seems that the summarizing dimension produced by the homogeneity analysis often failed to be collinear to the latent ability. Since the frequencies of some choices were low in some samples, the performance of this scoring method might be better in large samples; however, in such cases it might be more appropriate to use an item-response model.

The correlation-weighted scores passed our test quite well. When a positively discriminating distracter was present, they were the most valid score type, and in conditions with no such alternative they were about equally or marginally less valid than the number right scores. Besides the satisfactory validity evidence, these scores also had high reliability estimates and high unidimensionality measures. With an appropriate data entry, the correlation-weighted scores are simple to compute even with basic software like MS Excel. Because the number-right score is used as the criterion for determination of weights, the resulting weights should be closely related to the measured ability. Since they depend on bivariate relationships, they can be expected to be relatively stable in small samples, at least more than the homogeneity analysis weights or the multiple regression weights.

A question may be asked, whether MC items where more than one alternative is at least partially correct are appropriate at all. Of course, if the number-right scoring or even the guessing correction is adopted, we have to construct distracters that are undoubtedly incorrect. However, in scholastic tests this is easy to achieve only if the basic components of knowledge are measured, like the reproduction of simple facts. If the question involves higher levels of knowledge (like application, analysis, evaluation etc.), it may be unnatural to demand that all alternatives except one are completely incorrect. Similarly, when solving an ability item involves a combination of operations (for instance, simultaneous rules in matrix-like tests), it would be more natural to allow alternatives which correspond to the use of some, but not all necessary operations. The use of correlation-weighted scores might therefore make the process of item construction more natural.

It may seem strange that the ECV proportions were relatively low (about 50%), as there was a single underlying latent trait, so ECV proportion should be close to 100%. An important factor lowering ECV was a small sample size – because of the sampling error, various irrelevant and sample-dependent common factors emerge. As proven by Shapiro (1982), in any real set of 20 variables the minimum number of common factors is higher than 14. An additional factor may have been the distortion caused by the translation of a non-linear item-response model to a linear data analytic model. Admittedly, a linear common factor analysis may not be the optimal method for assessing the dimensionality of discrete item responses, however, the minimum rank factor analysis that we used is the only available latent trait method that produces a useful descriptive measure of unidimensionality.

Finally, some limitations of our study should be outlined. The latent structure of data was perfectly unidimensional, which is not a realistic condition. However, we believe than none of the methods we used is particularly affected by a multidimensional latent structure. Further, we have to stress again that only the lower bounds to reliability were compared rather than reliability in the strict sense. This problem could be overcome by an additional simulation of a retest, but in our opinion this would not be worth the effort, since the comparison of reliability was not our primary goal. It is also difficult to see why the GLBR values would be systematically lower for some particular scoring method compared to another.

It is important to note that we did not explicitly control the guessing proneness of our virtual examinees. The individual differences in guessing arose by chance. Had there been extreme individual differences systematically introduced, we could expect a somewhat better performance of the guessing correction scoring. However, in testing practice examiners usually attempt to make the individual differences in guessing as small as possible.

We did not attempt to estimate the stability of the weights obtained by the CW and the HA scoring, respectively. The stress of the study was on the use of scoring methods in a particular sample.

To conclude, our study pointed to the correlation-weighted scoring as a possibly useful method for an optimal summarization of information contained in responses to multiple choice items. It would be premature to recommend this type of scoring to be widely used at this stage, but we certainly think that psychometricians may well pay attention to correlation weights.

REFERENCES

Alnabhan, M. (2002). An empirical investigation of the effects of three methods of handling guessing and risk taking on the psychometric indices of a test. *Social Behavior and Personality, 30,* 645-652.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37,* 29-51.

Bock, R. D. (1997). The nominal categories model. In W. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 33-49). New York: Springer.

Brennan, R. L. (2006). *Educational measurement.* Westport, CT: ACE/Praeger

Burton, R. F. (2004). Multiple choice and true/false tests: reliability measures and some implications of negative marking. *Assessment & Evaluation in Higher Education, 29,* 585-595.

Burton, R. F. (2005). Multiple choice and true/false tests: myths and misapprehensions. *Assessment & Evaluation in Higher Education, 30,* 65-72.

Coombs, C. H. (1953). On the use of objective examinations. *Educational and Psychological Measurement, 13,* 308-310.

Cross, L. H., & Frary, R. B. (1977). An empirical test of Lord's theoretical results regarding formula scoring of multiple-choice tests. *Journal of Educational Measurement, 14,* 313-321.

De Gruijter, D. N. M., & van der Kamp, L. J. T. (2008). *Statistical test theory for the behavioral sciences.* Boca Raton, FL: Chapman&Hall/CRC.

Frary, R B. (1982). A simulation study of reliability and validity of multiple-choice test scores under six response-scoring modes. *Journal of Educational Statistics, 7,* 333-351.

Gibbons, J. D, Olkin, I., & Sobel, M. (1997). A subset selection technique for scoring items on a multiple choice test. *Psychometrika, 44,* 259-270.

Gifi, A. (1990). *Nonlinear multivariate analysis.* Chichester: Wiley.

Greenacre, M. (2007). *Correspondence analysis in practice* (2nd ed.) Boca Raton, FL: Champan&Hall/CRC.

Gulliksen, H. (1950). *Theory of mental tests.* New York: Wiley.

Hofstee, W. K. B, Ten Berge, J. M. F, & Hendriks, A. A. J. (1997). How to score questionnaires. *Personality and Individual Differences, 25,* 897-909.

Kansup, W., & Hakstian, A. R. (1975). A comparison of several methods of assessing partial knowledge in multiple-choice tests: I. Scoring procedures. *Journal of Educational Measurement, 12,* 219-230.

Lord, F. M. (1975). Formula scoring and number-right scoring. *Journal of Educational Measurement, 12,* 7-11.

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

MATLAB 5.3 [Computer software].(1998). Natick, MA: MathWorks.

McDonald, R. P. (1999). *Test theory: A unified treatment.* Mahwah, NJ: Lawrence Erlbaum.

Michailidis, G., & de Leeuw, J. (1998). The Gifi system of descriptive multivariate analysis. *Statistical Science, 13,* 307-336.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.).New York: McGraw-Hill.

Poizner, S. B., Nicewander, W. A., & Gettys, C. F. (1978). Alternative response and scoring methods for multiple-choice items: An empirical study of probabilistic and ordinal response modes. *Applied Psychological Measurement, 2,* 83-96.

Reid, F. (1977). An alternative scoring formula for multiple-choice and true-false tests. *Journal of Educational Research, 70,* 335-339.

Revuelta, J. (2005). An item response model for nominal data based on the rising selection ratios criterion. *Psychometrika, 70,* 305-324.

Shapiro, A. (1982). Rank-reducibility of a symmetric matrix and sampling theory of minimum trace factor analysis. *Psychometrika, 47,* 187-199.

Ten Berge, J. M. F. (1993). *Least squares optimization in multivariate analysis.* Leiden: DSWO Press.

Ten Berge, J.M.F., & Kiers, H.A.L. (1991). A numerical approach to the approximate and the exact minimum rank of a covariance matrix. *Psychometrika, 56,* 309-315.

Ten Berge, J.M.F., Snijders, T.A.B., & Zegers, F.E. (1981). Computational aspects of the greatest lower bound to

the reliability and constrained minimum trace factor analysis. *Psychometrika, 46,* 201-213.

Ten Berge, J.M.F., & Sočan, G. (2004). The greatest lower bound to the reliability of a test and the hypothesis of unidimensionality. *Psychometrika, 69,* 613-625.

Thissen, D., & Steinberg, L. (1984). A response model for multiple choice items. *Psychometrika, 49,* 501-519.

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika, 51,* 567-577.

Thissen, D., & Steinberg, L. (1997). A response model for multiple-choice items. In W. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 51-65). New York: Springer.

Thurstone, L. L. (1919). A scoring method for mental tests. *Psychological Bulletin, 16,* 235-240.

Wainer, H. (1976). Estimating coefficients in linear models: It don't make no nevermind. *Psychological Bulletin, 83,* 213-217.

Zimmerman, D. W., & Williams, R. H. (2003). A new look at the influence of guessing on the reliability of multiple-choice tests. *Applied Psychological Measurement, 27,* 357-371.

APPENDIX

Parameter values for the data generation

Condition	Parameter	Correct	I1	I2	I3	Omit
I. One distracter with positive discrimination, moderate guessing	*a*	0.90	0.80	-0.40	-0.50	-0.80
	c	1.32	-0.07	-0.83	-0.77	0.34
II. One distracter with positive discrimination, high guessing	*a*	0.90	0.80	-0.40	-0.50	-0.80
	c	1.34	0.19	0.09	-0.21	-1.41
III. No distracters with positive discrimination, moderate guessing	*a*	1.00	0.00	-0.30	-0.20	-0.50
	c	1.24	-0.03	-0.86	-0.80	0.44
IV. No distracters with positive discrimination, high guessing	*a*	1.00	0.00	-0.30	-0.20	-0.50
	c	1.54	0.71	0.11	0.11	-0.94

Note. I1-I3 = alternatives scored as incorrect. Symbols *a* and *c* are the same as in Bock (1972, 1997).