Metodološki zvezki, Vol. 12, No. 1, 2015, 25-43

Empirical Option Weights for Multiple-Choice Items: Interactions with Item Properties and Testing Design

Gregor Sočan[1]

Abstract

In scoring of a multiple-choice test, the number of correct answers does not use all information available from item responses. Scoring such tests by applying empirically determined weights to the chosen options should provide more information on examinees' knowledge and consequently produce more valid test scores. However, existing empirical evidence on this topic does not clearly support option weighting. To overcome the limitations of the previous studies, we performed a simulation study where we manipulated the instruction to examinees, discrimination structure of distractors, test length, and sample size. We compared validity and internal consistency of number-correct scores, corrected-for-guessing scores, two variants of correlation-weighted scores and homogeneity analysis scores. The results suggest that in certain conditions the correlation-weighted scores are notably more valid than the number-correct scores. On the other hand, homogeneity analysis cannot be recommended as a scoring method. The relative performance of scoring methods strongly depends on the instructions and on distractors' properties, and only to a lesser extent on sample size and test length.

1 Introduction

The multiple-choice format is a popular item format for ability and attainment tests, especially when a maximally objective or even an automated scoring is desired, or when the use of constructed-response items would be impractical. However, the relatively complex form of a multiple-choice item allows for

[1] Gregor Sočan, Department of psychology, University of Ljubljana, Aškerčeva c. 2, SI-1000 Ljubljana, Slovenia; gregor.socan@ff.uni-lj.si

different scoring techniques, which differ with regard to the information they take into account and the specific algorithm for transforming this information into a single numeric score. Since guessing can be an important factor contributing to the test score, the instructions for examinees and the scoring algorithm should not be sensitive to individual differences in the guessing-related attitudes and processes. This paper deals with the practical usefulness of scoring techniques, based on empirically derived weights, which are consistent with the classical test theory (thus excluding item response theory models). Furthermore, we shall limit our treatment to the most common form of items, where the examinee is required to choose one of the m offered alternatives, $m > 2$, excluding item formats like true/false, answer-until-correct, option elimination, confidence weighting etc.

1.1 Scoring based on the number of correct responses only

The simplest scoring rule is the number-correct (NC) rule: the score is determined as a number of correct responses. The most obvious advantage of the NC rule is its conceptual and computational simplicity. On the other hand, the test scores are contaminated with the examinees' tendency to guess, unless all examinees respond to all items. Moreover, the NC rule does not use all information available from the item response. In particular, when the examinee does not respond correctly, (s)he gets zero points regardless of the degree of incorrectness of the chosen distractor.

To eliminate the effect of individual differences in the guessing proneness, the correction for guessing (CG), also known as "formula scoring" can be used. This rule assigns different scores to incorrect responses and to omitted responses; typically, an incorrect response is scored with $-[1/(m-1)]$ points, and an omitted response is scored with 0 points[2]. Although the CG scoring could be treated as a simple case of differential response weighting, this conceptualization would imply that the differences between 1, 0 and $-[1/(m-1)]$ points reflect the differences in expected values of the respective examinees' knowledge – a position that would be disputable at best. It is therefore more meaningful to treat it as a modification of the NC scoring. In any case, the opinions about the correction for guessing have varied widely (for a recent review see Lesage, Valcke, and Sabbe, 2013). In short, the advocates have mostly stressed the higher measurement precision relative to the NC scores (Burton, 2004, 2005; Lord, 1975), which is achieved when some basic assumptions about the response process hold. On the other hand, the critics (for instance, Bar-Hillel, Budescu, and Attali, 2005; Budescu and Bo, 2015; Cross and Frary, 1977) pointed to introduction of biases related to the individual differences in risk aversion, accuracy of the subjective estimation of the probability of a correct response, and similar psychological factors. The issue of

[2] Alternative CG formulas and algorithms have been proposed; for an evaluation see Espinosa and Gardeazabal (2010, 2013).

correction for guessing is complicated by the finding that statistically equivalent variants of the CG scoring are not necessarily strategically equivalent when the risk aversion is taken into account (Espinosa and Gardeazabal, 2013), and by the fact that researchers working with similar quantitative models of examinees' responding to multiple choice items sometimes arrived to opposite conclusions about the optimal size of the penalty for a wrong answer (Budescu and Bo, 2015 vs. Espinosa and Gardeazabal, 2013). We shall not extend our discussion of the CG scoring, because it does not play the central role in this study; however, we should stress that various irrelevant psychological factors may significantly determine the psychometric properties of guessing-corrected scores.

1.2 Empirical option weighting

The previously discussed scoring rules disregard the information contained in the particular choice of an incorrect option. That is, if the examinee chooses an incorrect option, the item score does not depend on which option has been chosen. If the distractors differ in the level of incorrectness, which may often be the case, such scoring does not use all information contained in the item response. Therefore, taking account of a choice of a particular distractor should in principle increase reliability and validity of the examinee's score. One possible way of using this information is by using an item-response theory (IRT) model, modelling the relation between the latent trait and the response to each option. Such models include cases described by Bock (1997), Thissen and Steinberg (1997), Revuelta (2005), and García-Pérez and Frary (1991). However, if the sample size is not very large or when the assumptions of the IRT models are not satisfied – for instance, in areas like educational measurement or personality assessment the measured construct is often multidimensional – researchers may prefer to use classical models, based on linear combinations. A computational approach which seems particularly attractive is homogeneity analysis (HA; also known as dual scaling, optimal scaling, multiple correspondence analysis, and Guttman (1941) weighting). Homogeneity analysis transforms qualitative variables into quantitative variables by means of weighting the elements of the indicator matrix, corresponding to the item options.

In case of multiple choice items, the responses to each of k items are recorded as a categorical variable with m categories (or possibly $m + 1$ categories, if omissions are to be weighted as well). This variable is then recoded into m (or $m + 1$, respectively) indicator variables, taking the value of 1 if the corresponding option was selected by the examinee, and the value of 0 otherwise. Then, k vectors of weights are computed minimizing the discrepancy function

$$f(\mathbf{w}_1, \ldots, \mathbf{w}_k, \mathbf{h}) = \sum_{j=1}^{k} \left\| \mathbf{G}_j \mathbf{w}_j - \mathbf{h} \right\|^2 \quad (1)$$

where $\mathbf{G}_j$ is a matrix of indicator variables for item *j*, and **h** is a vector of weighted test scores, whereas the symbol $\|\mathbf{X}\|^2$ stands for the sum of squared elements of the matrix **X**. Since Equation 1 formally represents a multivariate regression model, we can interpret the score vector **h** as the best linear approximation of the quantified item responses in the least squares sense. That is, the score explains the maximum possible amount of quantified items' variances. Conversely, the quantified variables are maximally homogeneous, which means that they have minimum within-person variance. Although several uncorrelated score vectors can be determined, only the first one is interesting in our case.

From the psychometric viewpoint, a particularly attractive feature of HA is that its weights maximize the value of coefficient alpha for the total test score, calculated as the sum of the quantified variables. From the interpretational viewpoint, HA is attractive because of its close relation to the principal component analysis (PCA). HA can be understood as a variation of PCA for categorical variables; note that the first principal component also maximizes the explained variances of the analyzed variables and the coefficient alpha for their sum (for further details see Gifi, 1990, Greenacre, 2007, and Lord, 1958).

Two potential drawbacks of HA should be noted. As a statistical optimization technique, it is prone to chance capitalization, and it is not clear what sample sizes are needed for a reliable generalization of multiple-choice item weights. Furthermore, although it makes intuitive sense to expect a high correlation between the score vector and the measured knowledge, there is no guarantee for this to happen in a particular dataset. Fortunately, the seriousness of both potential problems can be assessed by means of simulation.

A compromise solution, sacrificing optimality to achieve greater stability, would consist of weighting the indicator variables with their point-biserial correlations with the total test score. The correlation weights (CW) should generally result in a smaller amount of quantified items' homogeneity, but on the other hand their use should reduce the scope of the above mentioned limitations of HA. Since the correlation weights are not aimed at maximizing homogeneity in the sample, they do not capitalize on chance; and because the total test score is always relatively highly correlated with knowledge, the weighted score cannot have a very low correlation with knowledge. Although not producing optimal scores, the correlation weights are appealing because they still give more importance to better discriminating options. Furthermore, the calculation of correlations with the total unweighted score represents the first step in Guttman's (1941) version of HA. Although this computational procedure for HA is now obsolete, this fact shows that CW can be understood as the first-step approximation to the HA solution – similarly as the classical item discrimination coefficients are the first-step approximations to the first principal component loadings.

Both approaches to the empirical option weighting were evaluated in several older studies, which mostly focused on the internal-consistency reliability and the

criterion validity of the weighted option scores in large samples. Davis and Fifer (1959) found that the cross-validated correlation weights, compared to the NC scores, increased reliability, but not validity. Reilly and Jackson (1973), Downey (1979), and Hendrickson (1971) reported that the HA scores, compared to the NC scores, were more reliable and less valid. Sabers and Gordon (1969) found no notable differences in either reliability or validity. Echternacht (1976), on the other hand, reported higher values of both reliability and validity coefficients for the HA scores compared to the NC scores.

Almost all studies found higher internal consistency of HA scores compared to NC scores; however, this should be expected due to the alpha-maximizing property of HA. Inconsistent results with regard to criterion validity are more puzzling. Echternacht (1976) simply attributed the inconsistency to the fact that different criterion variables implied different validity coefficients. Hendrickson (1971) conjectured that lower correlations with other variables were a consequence of a more homogeneous factor structure of HA items scores relative to the unweighted item scores.

Reilly and Jackson (1973) noted that the weights assigned to omitted responses were often very low in practice. Raffeld (1975) assigned a constant zero weight to omitted responses, which improved predictive validity of the weighted option scores. However, this solution does not seem completely satisfactory, because setting the value of a weight to zero does not have a clear rationale. Nevertheless, Raffeld's results are important because they highlight the issue of weighting the omitted responses.

The inconsistent results concerning validity of option-weighted scores eventually led Frary (1989) to conclude that "option weighting offers no consistent basis for improving the psychometric quality of test scores unless there is a problem with respect to internal consistency reliability." (p. 83). However, the preceding empirical studies had a common limitation: for validity assessment, they relied on the correlation with external criterion variable(s). A more definite conclusion could be reached if the correlation with the response-generating trait was determined. Sočan (2009) simulated test scores while manipulating the degree of incorrectness of distractors and the general level of guessing. The results predictably showed better homogeneity and reliability of both CW and HA scores, compared to the NC scores, while the validity (i.e., correlation with the latent trait) of the CW scores was either marginally worse or notably better than the validity of the NC scores, the difference being larger when there was a distractor with a positive discrimination (for instance, a partially correct alternative). However, Sočan's study had some limitations. First, only performance in small samples was studied, and the weights were neither cross-validated nor generalized to the population. Second, risk-aversion / guessing-proneness was not included as a person characteristic, which is not consistent with empirical evidence indicating the existence of guessing tendency as a stable personality-related trait (for instance, Brenk and Bucik, 1994; Dahlbäck, 1990; Slakter, 1967).

The problem of this study was to assess the psychometric quality (especially construct validity) of the correlation-weighted (CW) and homogeneity analysis (HA) scores compared to the number-correct scores (NC) and "guessing-corrected" (CG) scores. For all types of weighted option scores, we aimed to determine the generalizability of sample weights to the population with relation to sample size, test length, and distractor characteristics.

2 Method

2.1 Experimental treatments

The following four experimental conditions were manipulated in a crossed design.
1. Test instructions. In the first level of the factor ("forced choice"), simulated examinees were forced to choose one of the alternatives, regardless of their confidence in the correctness of their choice. In the second level ("omissions allowed"), they were instructed to respond only when being reasonably confident that the selected alternative was correct.
2. Distractor properties. Each test consisted of multiple-choice items with three alternatives. This format was chosen both for simplicity and according to recommendations from the meta-analysis by Rodriguez (2005). The alternatives were characterized by its discriminating power, defined as the correlation with the measured trait (denoted by $r_{j(a)\theta}$). The discrimination parameter for the correct option was always $r_{j(c)\theta} = .40$. The discrimination parameters for the two distractors were different in the two levels of the factor, namely:
   1. "both negative": $r_{j(1)\theta} = -.20$ and $r_{j(2)\theta} = -.10$;
   2. "one positive": $r_{j(1)\theta} = -.20$ and $r_{j(2)\theta} = .10$.

The first factor level corresponds to the case where both distractors are clearly incorrect, while the second factor level corresponds to the case with one partially correct distractor. In both cases, all items in the test had equal discrimination structure. The values of parameters were set according to the results of preliminary analyses on real data (not shown here).
3. Test length. Two test lengths were used, with the number of items $k = 15$ and 50, respectively.
4. Sample size. Three sample size levels were used: $n = 100$, 200 and 500, respectively.

2.2 Simulation design

In the first step, a population was defined as a group of one million examinees, characterized by two uncorrelated latent traits: knowledge (θ~N(0,1)) and risk-aversion (γ~N(0,0.5)). The variance of risk-aversion was smaller than the variance of knowledge to prevent the presence of examinees with extreme guessing/omitting behavior (see the description of the response model below). Responses were generated for each of the eight combinations of factors 1-3.

We used the underlying variables approach to the response generation. For each examinee, we computed his/her response propensity for each option:

$$x_{ij(a)}^{*} = \theta_i r_{j(a)\theta} + e_{ij(a)}, \quad (2)$$

where $e_{ij(a)}$ was a random error component, $e_{ij(a)}$~N(0,$(1-r^2_{j(a)\theta})^{1/2}$).

In the "forced choice" condition, the response was simply determined as the option with the largest response propensity. In the "omissions allowed" condition, an examinee responded only when $\max(x_{ij(a)}^{*}) > \gamma_i$, and omitted the response otherwise. Therefore, our response model assumes that an examinee finds the option which seems to be correct most likely, but omits the response (if allowed to do so) if his/her confidence in the correctness of the choice does not exceed the personal confidence criterion (which is higher for more risk-averse persons and vice versa).

When all responses were generated, 10000 random samples of each size were drawn without replacement from each of the population response matrices. Altogether, 3×8×10000 = 240000 samples were processed. In each sample, both correlation weights and homogeneity weights were determined. Two versions of correlation weights were calculated. In both cases the weights were calculated as Pearson correlations between the indicator variable corresponding to an option (including, where relevant, an omission) and the test score. The first version (CW_{NC}) used the number-correct score as the criterion variable, and the second version (CW_{CG}) used the score corrected for guessing. The HA weights were calculated using the closed form solution for homogeneity analysis (ten Berge, 1993, pp. 66-67). For the CG scoring, the standard penalty of -[1/(m-1)] = -0.5 points for an incorrect response was used. Five sets of scores were thus obtained for each sample, and the sample values of criterion variables were calculated. Then, the CW and HA weights obtained in the sample were applied to the population responses, and the values of criterion variables were calculated again. Table A in the appendix presents percentages of the choices and both validities and coefficients alpha for the number-correct score in various conditions.

All simulations were performed using MATLAB R2012b software. The MATLAB codes used for the simulations are available as supplementary material from the website of Metodološki zvezki (http://www.stat-d.si/mz/Articles.html).

3 Results

We shall first compare the four scoring rules with respect to validity and internal consistency of the obtained scores. After that, we shall compare both instruction conditions with respect to the validity of scores. We present no inferential statistics: because of large sample sizes and due to the use of the repeated-measures design, the statistical power was very high, and even some practically irrelevant effects reached statistical significance. All discussed results were statistically significant ($p < .001$).

3.1 Validity

The most important question addressed in this study is validity of scores obtained by different scoring rules. Since the responses were simulated, it was possible to assess the construct validity directly as the Pearson correlation between the test score and the latent knowledge. Table 1 presents validity increments or losses, respectively, i.e. differences between mean validity coefficients of the number-correct score and mean validity coefficients of each of the three remaining scores in various conditions. A positive difference indicates that a scoring rule produces more valid scores than the number-correct rule. The values in the left part of the table are related to the sample validities, while the values in the right part of the table are computed from the generalized validities; that is, they are based on population scores computed with sample weights. Since the CG weights are not empirically estimated, the respective population validities are independent of sample size.

Clearly, the instruction for examinees was a major determinant of the pattern of validity increments. When examinees could omit the response, both the CG scores and the CW_{CG} scores were invariably more valid than the NC scores. As expected, when there was a positively discriminating distractor, option weighting resulted in a slightly higher increment than the simple correction for guessing. On the other hand, when both distractors had negative discrimination power, the guessing correction produced higher increments than option weights, unless the sample size was at least 500. For both scoring rules, the increments were higher for longer tests. The remaining two weighting schemes produced scores less valid than the number-correct scores. In case of the HA scoring, the differences were especially large; in fact, the average validity of the HA scores was lower than .40 in all 12 conditions related to the "omissions allowed" instruction. The differences for the CW_{NC} scores were much smaller, but they were still of notable size and consistently negative. The increments and losses, respectively, of all four scoring methods were larger in absolute value when the test was longer.

In the forced choice condition, the increments for the CG scoring are not presented, because the CG scores are linearly related to the NC scores and are therefore equally valid. Consequently, the CW_{CG} rule and the CW_{NC} rule are also equivalent.

Table 1: Validity increments/losses relative to the NC scoring

				Sample				Population			
Ins.	*k*	$r_{j\theta} > 0$	*n*	CG	CW_{NC}	CW_{CG}	HA	CG	CW_{NC}	CW_{CG}	HA
FC	15	one	100	-	**.02**	--	-.02	-	**.02**	--	-.01
			200	-	**.03**	--	**.02**	-	**.03**	--	**.02**
			500	-	**.03**	--	**.04**	-	**.03**	--	**.04**
		none	100	-	-.01	--	-.04	-	-.01	--	-.04
			200	-	.00	--	-.02	-	.00	--	-.02
			500	-	**.00**	--	.00	-	**.00**	--	.00
	50	one	100	-	**.01**	--	**.01**	-	**.01**	--	**.01**
			200	-	**.02**	--	**.02**	-	**.02**	--	**.02**
			500	-	**.02**	--	**.02**	-	**.02**	--	**.02**
		none	100	-	-.01	--	-.01	-	-.01	--	-.01
			200	-	.00	--	.00	-	.00	--	.00
			500	-	**.00**	--	.00	-	**.00**	--	.00
OA	15	one	100	**.02**	-.02	**.04**	-.42	↑	-.02	**.03**	-.44
			200	**.02**	-.01	**.05**	-.43	**.02**	-.02	**.05**	-.43
			500	**.02**	-.01	**.05**	-.43	↓	-.01	**.05**	-.43
		none	100	**.04**	-.03	**.02**	-.52	↑	-.03	**.02**	-.58
			200	**.04**	-.02	**.03**	-.57	**.04**	-.03	**.03**	-.60
			500	**.04**	-.02	**.04**	-.59	↓	-.02	**.04**	-.60
	50	one	100	**.05**	-.06	**.06**	-.52	↑	-.08	**.05**	-.54
			200	**.05**	-.07	**.06**	-.53	**.05**	-.08	**.06**	-.53
			500	**.05**	-.07	**.06**	-.53	↓	-.07	**.06**	-.53
		none	100	**.05**	-.05	**.04**	-.64	↑	-.06	**.04**	-.69
			200	**.05**	-.05	**.05**	-.67	**.05**	-.05	**.04**	-.70
			500	**.05**	-.05	**.05**	-.69	↓	-.05	**.05**	-.69

Positive values are in boldface. Ins. = instruction, FC = forced choice, OA = omissions allowed, *k* = number of test items, *n* = sample size, $r_{j\theta} > 0$ = number of distractors with positive discrimination power, CG = correction for guessing, CW_{NC} = correlation weights (number-correct score as the criterion), CW_{CG} = correlation weights (guessing-corrected score as the criterion), HA = homogeneity analysis weights, - not relevant because both rules are equivalent, -- not reported because it is equal to the CW_{NC} value, ↑ equal to the value below, ↓ equal to the value above.

With the forced choice instruction, the average validities of the weighted option scores were generally higher than the validities of the NC scores when one distractor had a positive discrimination. For the CW scores, the validity increment was slightly higher for the shorter test than for the longer test. When both

distractors discriminated negatively, the differences were close to zero, except when the sample size was 100; in this case, the weighted option scores were less valid than the NC scores, and we may conjecture that the weighted option scores would be more valid than the NC scores in larger samples (i.e., larger than 500). In general, validity increments of the weighted option scores were larger in larger samples, however, the effect of the sample size was slim.

We can also note two general observations. First, the pattern of the sample differences was very similar to the pattern of the population differences. Therefore, for the evaluation of validity of scores obtained by various scoring methods, it does not matter much whether only a particular sample is of interest, or the generalization to the population is desired. Second, the average validity of (both variants of) CW scores was at least as high as the average validity of HA scores in almost all conditions, and it was much higher when it was possible to omit the response.

3.2 Internal consistency

In the broad sense, internal consistency is related to the homogeneity of a group of measures, in this case test items. We evaluated two aspects: internal consistency reliability and the amount of items' variance, explained by the first principal component. For the former, we used coefficient alpha, which should be an accurate estimate of reliability, since the items were constructed as parallel measurements. The values of coefficient alpha were calculated from the covariance matrices of the item scores. For the latter, we performed principal component analysis on the inter-item correlation matrix. We do not report sample results; it is a mathematical necessity that the HA scores are more internally consistent than the NC scores, and since the correlation weights are approximations to the HA weights, the same can be expected for both variants of CW scores. Indeed, all weighted option scores had higher sample values of both coefficient alpha and the explained variance than the NC scores.

However, when weights are cross-validated or generalized to the population, respectively, the superiority of weighted option scores is not guaranteed any longer. Table 2 presents internal consistency indicators for the scores, obtained by applying sample weights to the population response matrix. As in the previous section, the increments/losses relative to the NC scoring are presented. First, we can note a general – and, of course, expected - pattern of increasing internal consistency increments with sample size – that is, when the weights were obtained in larger samples, they generalized to the population better.

In all "omissions allowed" conditions, internal consistency of the HA scores was notably better, and internal consistency of the CG scores was somewhat worse in comparison to the NC scores. The behavior of CW scores was more complex. The CW_{NC} scoring resulted in positive increments where one distractor had a

positive discrimination; otherwise, the increment was positive only when the sample size was relatively large. On the other hand, the increments for the CW_{CG} scores were positive only when the test was long, the sample size was 500 and there was a positively discriminating distractor.

Table 2: Internal consistency increments/losses relative to the NC scoring

				α				Var_{PC1}			
Ins.	*k*	$r_{j\theta} > 0$	*n*	CG	CW_{NC}	CW_{CG}	HA	CG	CW_{NC}	CW_{CG}	HA
FC	15	one	100	-	**.03**	--	-.03	-	**1.03**	--	**0.23**
			200	-	**.04**	--	**.03**	-	**1.28**	--	**1.12**
			500	-	**.05**	--	**.05**	-	**1.42**	--	**1.64**
		none	100	-	-.02	--	-.07	-	-0.34	--	-1.40
			200	-	-.01	--	-.03	-	-0.07	--	-0.54
			500	-	.00	--	-.01	-	**0.08**	--	-0.08
	50	one	100	-	**.02**	--	**.01**	-	**1.20**	--	**0.90**
			200	-	**.03**	--	**.03**	-	**1.59**	--	**1.48**
			500	-	**.04**	--	**.04**	-	**1.82**	--	**1.83**
		none	100	-	-.02	--	-.02	-	-0.53	--	-0.88
			200	-	-.01	--	-.01	-	-0.15	--	-0.31
			500	-	.00	--	.00	-	**0.07**	--	**0.01**
OA	15	one	100	-.07	**.05**	-.03	**.15**	↑	**2.13**	-.57	**7.33**
			200	-.07	**.06**	-.01	**.17**	-1.82	**2.54**	-.22	**8.51**
			500	-.07	**.07**	.00	**.18**	↓	**2.78**	-.02	**9.15**
		none	100	-.03	-.01	-.05	**.07**	↑	-.19	-1.55	**4.12**
			200	-.03	.00	-.04	**.09**	-1.13	**.20**	-1.18	**5.41**
			500	-.03	**.01**	-.03	**.11**	↓	**.42**	-.97	**6.10**
	50	one	100	-.04	**.03**	-.02	**.08**	↑	**3.04**	-.25	**8.27**
			200	-.04	**.04**	.00	**.09**	-1.92	**3.68**	**.30**	**9.18**
			500	-.04	**.05**	**.01**	**.09**	↓	**4.07**	**.61**	**9.70**
		none	100	-.02	-.01	-.03	**.04**	↑	-.16	-1.82	**5.01**
			200	-.02	**.00**	-.02	**.05**	-1.18	**.41**	-1.30	**5.99**
			500	-.02	**.01**	-.01	**.05**	↓	**.75**	-.98	**6.53**

Positive values are in boldface. α = coefficient alpha of internal consistency reliability, Var_{PC1} = percentage of variance explained by the first principal component, *k* = number of test items, $r_{j\theta} > 0$ = number of distractors with positive discrimination power, *n* = sample size, CG = correction for guessing, CW_{NC} = correlation weights (number-correct score as criterion), CW_{CG} = correlation weights (guessing-corrected score as criterion), HA = homogeneity analysis weights, - not relevant because both rules are equivalent, ↑ equal to the value below, ↓ equal to the value above.

In the "forced choice" conditions, the weighted option scores were in general more internally consistent in cases with a positively discriminating distractor; otherwise, the internal consistency of weighted option scores was comparable to the internal consistency of the NC scores when the weights were obtained in

samples of size 500, and somewhat smaller when obtained in smaller samples. Therefore, when examinees are required to choose an option, using empirical option weights might significantly increase the population internal consistency only in samples of size considerably larger than 500. The performance of the HA scoring was generally comparable or even slightly worse than the performance of the CW scoring.

3.3 Effect of instructions on validity

Table 3: Validity increments of the "forced choice" over the "omissions allowed" instructions

			Sample					Population				
k	$r_{j\theta} > 0$	n	NC	CG	CW_{NC}	CW_{CG}	HA	NC	CG	CW_{NC}	CW_{CG}	HA
15	one	100	.03	.01	.07	.01	.43	↑	↑	.07	.01	.46
		200	.03	.01	.07	.01	.48	.03	.01	.07	.01	.48
		500	.03	.01	.07	.01	.49	↓	↓	.07	.01	.50
	none	100	.05	.01	.07	.01	.53	↑	↑	.07	.02	.58
		200	.05	.01	.07	.01	.60	.05	.01	.07	.01	.63
		500	.05	.01	.07	.01	.63	↓	↓	.07	.01	.64
50	one	100	.06	.01	.14	.02	.59	↑	↑	.16	.03	.61
		200	.06	.01	.15	.02	.60	.06	.01	.16	.02	.61
		500	.06	.01	.15	.02	.61	↓	↓	.16	.02	.61
	none	100	.06	.01	.11	.01	.69	↑	↑	.11	.02	.74
		200	.06	.01	.11	.01	.73	.06	.01	.11	.02	.76
		500	.06	.01	.11	.01	.75	↓	↓	.11	.02	.76

k = number of test items, $r_{j\theta} > 0$ = number of distractors with positive discrimination power, n = sample size, NC = number-correct, CG = correction for guessing, CW_{NC} = correlation weights (number-correct score as criterion), CW_{CG} = correlation weights (guessing-corrected score as criterion), HA = homogeneity analysis weights, ↑ equal to the value below, ↓ equal to the value above.

A question can be posed whether one type of instruction generally results in a higher average validity, and whether this effect is moderated by the choice of a scoring method. Table 3 presents the average differences between validity coefficients in the "forced choice" condition and the same coefficients in the "omissions allowed" condition. The presented results show that the instructions for examinees strongly determine the performance of various scoring methods. All differences were positive: the average validity in a specified test design was always higher in the "forced choice" than in "omissions allowed" condition. The sizes of the average differences were very similar for both sample and population validities. The differences were in general larger for longer tests, whilst the effect of sample size was negligible. Although the instruction effect was present in all

types of scores, the average size of the difference varied: it was quite small for the CG and CW_{CG} scores, and very large for the HA scores; in the latter case, the extreme superiority of the "forced choice" instruction was related to the very poor validity of the HA scores in the "omissions allowed" condition, as reported in section 3.1.

4 Discussion

We see the main contribution of this study in elucidating the interactions between various aspects of testing design and the performance of different scoring methods. If examinees are instructed to attempt every item, the expected validity is higher than if they are instructed to avoid guessing. This fact reveals that individual differences in guessing behavior are generally more detrimental to measurement quality than the guessing itself. Although the instruction effect is present generally, its size depends on the scoring method and - to a smaller extent – on the test length. The effect of test length can be attributed to the law of large numbers: with a larger number of items, the proportion of "lucky guesses" converges to its expectation. Lengthening the test therefore reduces the effect of luck, but does not affect the effect of personality factors of guessing behavior.

When examinees answer all items (as in the "forced-choice" condition), the number-correct scoring should be preferred to the option-weighting scoring if all distractors have negative discriminations of similar size; option-weighted scores seem to be more valid only in samples much larger than 500. On the other hand, when partially correct distractors (with a small positive discrimination power) are included, option weighting increases validity compared to the number-correct scoring, even in small samples (like n=100).

When guessing is discouraged and examinees omit some items (the "omissions allowed" condition), the correction for guessing and correlation weighting (based on the guessing-corrected scores) should be preferred methods. The CW_{CG} scoring may be the method of choice if there is a positively discriminating distractor; when all distractors have negative discrimination, the CG scores are more valid than the CW_{CG} scores if the sample size is less than about 500. Note than the CW_{CG} rather CW_{NC} weights should be used with this instruction. Because the CW scoring uses more information than the CG scoring, we speculate that it may be less sensitive to personality factors like subjective utility (as discussed in Budescu and Bo, 2015, and Espinosa and Gardeazabal, 2013). When both types of scores have similar validity and reliability, it may be thus safer to use the CW_{CG} scores.

According to our results, homogeneity analysis cannot be recommended as a scoring method. The HA scores were never notably more valid than the correlation-weighted scores, and they were substantially less valid when examinees could omit items. Obviously, the relatively high internal consistency of

the HA scores in the "omissions allowed" condition does not imply a high validity, but only reflects the contamination of true scores with the non-cognitive variables determining the guessing level. In this condition, the classical true score is not a faithful representation of knowledge, but rather a conglomerate of knowledge and risk-aversion. Therefore, the HA score may be a highly reliable measure of a conceptually and practically irrelevant quantity. Clearly, reliability is not a relevant test characteristic in such circumstances. As a collateral contribution, our results thus illustrate the danger of relying solely on (internal consistency) reliability in evaluation of measurement quality of cognitive tests.

The unsatisfactory behaviour of the HA weights may be surprising. It can be partly attributed to the fact that the HA algorithm does not assure a high correlation between the weighted score and the number correct score, which is especially problematic when there are other significant determinants of the examinee's response (in particular, the guessing proneness). The performance of the HA weights was better when the guessing proneness was controlled (by forcing all examinees to answer all items); although their validity was comparable to the validity of their approximations (i.e., the CW weights) in our sample size range, we may speculate that the HA weights might be superior in terms of validity in very large samples (for instance, in large standardized educational assessments), provided that both empirical scoring key and the forced-choice instruction would be considered acceptable.

In the "omissions allowed" conditions, the CW_{NC} and HA weights on one hand and the CW_{CG} weights on the other hand performed markedly different with respect to validity. These differences should be attributed to the differences in treatment of omitted responses. The number-correct score, controlled for the level of knowledge, is higher for examinees with a lower level of risk-aversion (that is, for examinees who attempt more items). As a consequence, a CW_{NC} weight for an omitted response is negative even if the actual decision between responding and omitting is completely unrelated to the level of knowledge. Indeed, the inspection of sample weights for omissions (not reported here) showed that the values of CW_{NC} weights were typically close to the weights for the (negatively discriminating) distractor(s), while the corresponding CW_{CG} weights were much closer to zero (cf. Reilly and Jackson, 1973). With the homogeneity analysis weights, the problem is essentially the same, but is aggravated due to a lack of a mechanism that would ensure at least approximate collinearity with knowledge. Using CW_{CG} weights removes the negative bias from the omission weights, however, these weights still reflect only one determinant of a response omission (i.e., knowledge), and disregard the personality-related determinants (risk aversion, subjective utility of a score, self-confidence and so on).

Increasing sample size seems to improve the performance of correlation-weighted scores. However, the sample size effect was quite small overall: weights determined in samples of 100 persons did not generalize substantially less well

than the weights determined in samples of 500 persons, especially when the test was long (k=50).

It should be noted that the patterns rather than particular values of the reported effects are to be taken as findings of this study. The values of the increments/losses depend on particular experimental settings, which are to a certain extent arbitrary. For instance, we can reasonably expect that using much easier items would make option-weighting less relevant because of a smaller percentage of incorrect responses. On the other hand, distractors with more heterogeneous discriminations would probably make the performance of the CW scores more favorable relative to the NC scores. Increasing sample size also seems to improve the performance of the CW scoring. However, a researcher who has collected a very large sample may first try to fit a more sophisticated item-response theory scoring model.

The research of scoring models for multiple-choice items is complicated by the lack of formalized cognitive models explaining the item response process. The existing models (for instance, Budescu and Bo, 2015, Espinosa and Gardeazabal, 2013) are mainly focused on guessing behavior, and are not useful in the context of weighting item options. Our study was based on very simple response model, which predicts examinee's decisions with just two person parameters (knowledge and risk-aversion). The model rests on two assumptions:

1. Both knowledge and risk-aversion are stable personal properties, which do not change during the testing process.
2. An examinee's response to a multiple-choice item is based on a comparison of plausibility of offered alternatives. When guessing is discouraged, the examinee omits the response if none of the option plausibility estimates exceeds the subjective criterion, determined by his/her level of risk-aversion.

While it seems safe to generally accept the first assumption, the second assumption is not so evidently generalizable to all examinees and testing contexts, and should be tested empirically in the future.

Empirical weighting should not be used with negatively discriminating items. Although the resulting negative weights for correct responses would generally increase both validity and internal consistency of the empirically weighted scores in comparison to the NC and CG scores, respectively, such scoring would be difficult to justify to test takers, developers and administrators. Of course, the presence of negatively discriminating items is problematic regardless of which scoring method has been chosen.

Furthermore, the weighting techniques assume invariance of item-option discrimination parameters across examinee subgroups. In real applications, this assumption may sometimes be violated. For instance, in intelligence testing, the discrimination parameters might be different for groups of examinees using different problem-solving strategies. In educational testing, using different textbooks or being exposed to different teaching methods might also cause

differences in discrimination parameters. Fortunately, the validity of the invariance assumption can be empirically verified, if the potentially critical subgroups can be identified.

5 Conclusions

Our results confirm Nunnally and Bernstein's (1994, p. 346) recommendation to instruct examinees to attempt every item. This instruction should not be questionable in psychological testing, especially when applying computerized tests, which can prevent possible accidental omissions. Correlation weights can be used to maximize the score validity if the distractors differ in their degree of incorrectness. In educational testing[3], some test administrators may not be comfortable with forcing the students to guess when they really do not recognize the correct answer. Using correlation weights, based on the corrected-for-guessing sum score, can be recommended in this case, especially if partially correct distractors have been used and the sample size is not too small. However, because omissions depend on both knowledge and risk-aversion, the scoring schemes studied here do not provide optimal scores for the omitted responses. Consequently, the validity of scores obtained with this instruction is lower compared to the "forced-choice" scores. Development of scoring models which would incorporate information on examinees' risk-aversion and other relevant personal characteristics, remains a task for future research.

References

[1] Bock, R.D. (1997): The nominal categories model. In W. van der Linden and R. K. Hambleton (Eds): *Handbook of Modern Item Response Theory*, 33-49. New York: Springer.

[2] Brenk, K. and Bucik, V. (1994): Guessing of answers in objective tests, general mental ability and personality traits according to 16-PF questionnaire. *Review of Psychology*, **1**, 11-20.

[3] Budescu, D.V. and Bo, Y. (2015): Analyzing test-taking behavior: Decision theory meets psychometric theory. *Psychometrika*, **80**, 1105-1122.

[3] It should be stressed that the methods discussed here are optimal in the context of the normative measurement, where the individual differences in knowledge are of main interest. Criterion-referenced measurement (for instance, a typical classroom assessment of knowledge) may call for different approaches to scoring.

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

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

[6] Bar-Hillel, M., Budescu, D., and Attali, Y. (2005): Scoring and keying multiple choice tests: A case study in irrationality. *Mind and Society*, **4**, 3-12.

[7] Cross, L.H. and 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.

[8] Dahlbäck, O. (1990): Personality and risk-taking. *Personality and Individual Differences*, **11**, 1235–1242.

[9] Davis, F.B. and Fifer, G. (1959): The effect on test reliability and validity of scoring aptitude and achievement tests with weights for every choice. *Educational and Psychological Measurement*, **19**, 159-170.

[10] Downey, R.G. (1979): Item-option weighting of achievement tests: Comparative study of methods. *Applied Psychological Measurement*, **3**, 453-461.

[11] Espinosa, M.P and Gardeazabal, J. (2010): Optimal correction for guessing in multiple-choice tests. *Journal of Mathematical Psychology*, **54**, 415-425.

[12] Espinosa, M.P. and Gardeazabal, J. (2013): Do students behave rationally in multiple choice tests? Evidence from a field experiment. *Journal of Economics and Management*, **9**, 107-135.

[13] Frary, R.B. (1989): Partial-credit scoring methods for multiple-choice tests. *Applied Measurement in Education*, **2**, 79-96.

[14] García-Pérez, M.A. and Frary, R.B. (1991): Finite state polynomic item characteristic curves. *British Journal of Mathematical and Statistical Psychology*, **44**, 45-73.

[15] Gifi, A. (1990): *Nonlinear Multivariate Analysis.* Chichester: Wiley.

[16] Greenacre, M. (2007): *Correspondence Analysis in Practice* (2nd ed.). Boca Raton, FL: Chapman and Hall/CRC.

[17] Guttman, L. (1941): The quantification of a class of attributes: A theory and method of scale construction. In P. Horst (Ed): *The Prediction of Personal Adjustment*, 321-345. New York: Social Science Research Council.

[18] Hendrickson, G.F. (1975): The effect of differential option weighting on multiple-choice objective tests. *Journal of Educational Measurement*, **8**, 291-296.

[19] Lesage, E., Valcke, M., and Sabbe, E. (2013): Scoring methods for multiple choice assessment in higher education: Is it still a matter of number right

scoring or negative marking? *Studies in Educational Evaluation*, **39**, 188–193.

[20] Lord, F.M. (1958): Some relations between Guttman's principal components of scale analysis and other psychometric theory. *Psychometrika*, **23**, 291-296.

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

[22] MATLAB R2012B [Computer software]. Natick, MA: The MathWorks.

[23] Nunnally, J. and Bernstein, I. (1994): *Psychometric Theory* (3rd ed.). New York, NY: McGraw-Hill.

[24] Raffeld, P. (1975): The effects of Guttman weights on the reliability and predictive validity of objective Tests when omissions are not differentially weighted. *Journal of Educational Measurement*, **12**, 179-185.

[25] Reilly, R.R. and Jackson, R. (1973): Effects of empirical option weighting on reliability and validity of an academic aptitude test. *Journal of Educational Measurement*, **10**, 185-194.

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

[27] Rodriguez, M.C. (2005): Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice*, **24**, 3–13.

[28] Sabers, D.L and Gordon, W. (1969): The effect of differential weighting of individual item responses on the predictive validity and reliability of an aptitude test. *Journal of Educational Measurement*, **6**, 93-96.

[29] Slakter, M.J. (1967): Risk taking on objective examinations. *American Educational Research Journal*, **4**, 31-43.

[30] Sočan, G. (2009): Scoring of multiple choice items by means of internal linear weighting. *Review of Psychology*, **16**, 77-85.

[31] Ten Berge, J.M.F. (1993): *Least Squares Optimization in Multivariate Analysis.* Leiden: DSWO Press. Retrieved from http://www.ppsw.rug.nl/~kiers/leastsquaresbook.pdf

[32] Thissen, D. and Steinberg, L. (1984): A response model for multiple choice items. *Psychometrika*, **49**, 501-519.

Appendix

Table A: Percentages of choices of various alternatives and metric properties of the number-correct score

			Response						
Instruct.	k	Distractors	O	C	W1	W2		$r_{\theta.NC}$	α_{NC}
Forced choice	15	one positive	n/a	33.4	34.2	32.5		.73	.53
		both negative	n/a	34.1	33.2	32.7		.79	.63
	50	one positive	n/a	33.4	34.1	32.5		.89	.79
		both negative	n/a	34.0	33.2	32.7		.92	.85
		Average	n/a	33.7	33.7	32.6		.83	.70
Omissions allowed	15	one positive	16.9	27.7	28.3	27.1		.70	.59
		both negative	16.7	28.3	27.7	27.3		.74	.66
	50	one positive	16.9	27.7	28.3	27.1		.83	.83
		both negative	16.7	28.3	27.7	27.3		.85	.86
		Average	16.8	28.0	28.0	27.2		.78	.74

k = number of items, O = omit, C = correct, W = wrong, $r_{\theta.NC}$ = validity of the number-correct score in the population, α_{NC} = coefficient alpha for the number-correct score in the population.