Adv in Health Sci Educ
https://doi.org/10.1007/s10459-017-9805-y

Multiple true–false items: a comparison of scoring algorithms

Felicitas-Maria Lahner[1] · Andrea Carolin Lörwald[1] · Daniel Bauer[2] · Zineb Miriam Nouns[1] · René Krebs[1] · Sissel Guttormsen[3] · Martin R. Fischer[4] · Sören Huwendiek[1]

Received: 27 October 2016 / Accepted: 26 November 2017

Abstract Multiple true–false (MTF) items are a widely used supplement to the commonly used single-best answer (Type A) multiple choice format. However, an optimal scoring algorithm for MTF items has not yet been established, as existing studies yielded conflicting results. Therefore, this study analyzes two questions: What is the optimal scoring algorithm for MTF items regarding reliability, difficulty index and item discrimination? How do the psychometric characteristics of different scoring algorithms compare to those of Type A questions used in the same exams? We used data from 37 medical exams conducted in 2015 (998 MTF and 2163 Type A items overall). Using repeated measures analyses of variance (rANOVA), we compared reliability, difficulty and item discrimination of different scoring algorithms for MTF with four answer options and Type A. Scoring algorithms for MTF were dichotomous scoring (DS) and two partial credit scoring algorithms, PS_{50} where examinees receive half a point if more than half of true/false ratings were marked correctly and one point if all were marked correctly, and $PS_{1/n}$ where examinees receive a quarter of a point for every correct true/false rating. The two partial scoring algorithms showed significantly higher reliabilities ($\alpha_{PS1/n} = 0.75$; $\alpha_{PS50} = 0.75$; $\alpha_{DS} = 0.70$, $\alpha_{A} = 0.72$), which corresponds to fewer items needed for a reliability of 0.8 ($n_{PS1/n} = 74$; $n_{PS50} = 75$; $n_{DS} = 103$, $n_{A} = 87$), and higher discrimination indices ($r_{PS1/n} = 0.33$; $r_{PS50} = 0.33$; $r_{DS} = 0.30$; $r_{A} = 0.28$) than dichotomous scoring and Type A. Items scored with DS tend to be difficult ($p_{DS} = 0.50$), whereas items scored with $PS_{1/n}$ become easy ($p_{PS1/n} = 0.82$). PS_{50} and Type A cover the whole range, from easy to difficult items ($p_{PS50} = 0.66$; $p_{A} = 0.73$). Partial credit scoring leads to better

✉ Felicitas-Maria Lahner
felicitas-maria.lahner@iml.unibe.ch

[1] Department of Assessment and Evaluation (AAE), Institute of Medical Education, University of Bern, Konsumstr 13, 3010 Bern, Switzerland

[2] Department of Education and Media, Institute of Medical Education, University of Bern, Bern, Switzerland

[3] Institute of Medical Education, University of Bern, Bern, Switzerland

[4] Institute for Medical Education, University Hospital, LMU, Munich, Germany

Published online: 30 November 2017

psychometric results than dichotomous scoring. PS_{50} covers the range from easy to difficult items better than $PS_{1/n}$. Therefore, for scoring MTF, we suggest using PS_{50}.

Keywords Assessment · Medical education · Multiple choice · Multiple true–false · Scoring · Undergraduates

Background

Written exams consisting of multiple true–false (MTF) items are "potentially one of the most useful of objective test types" (Cronbach 1939, p. 628).

MTF is a type of multiple choice item in which examinees have to rate every answer option independently as true or false in the context established by the item stem (Case and Swanson 2002; Cronbach 1941; Krebs 2004). In contrast to the widespread single-best answer items (Type A), more than one answer can thus be correct (Case and Swanson 2002). This is useful in assessment, for instance when presenting a problem with more than one correct solution (Krebs 2004) while simultaneously avoiding negatively worded stems, which are viewed as a flaw in item construction (Haladyna et al. 2002). In contrast to Pick-N, in which the number of true answers sought is commonly stated, the number of true answers is not revealed in MTF, thus minimizing the effect of test-wiseness and enhancing validity. While in Multiple Mark (MM) items, the examinee has to decide which answers are true, in MTF, the examinee additionally has to decide which answers are false (Tarasowa and Auer 2013). In this regard, MTF enables a distinction between an examinee making a wrong choice (wrong box checked) or not making a choice at all, for whatever reason (no box checked) (Gross 1982).

Interestingly, the US National Board of Medical Examiners (NBME) recommends against the use of MTF, considering them to be more commonly flawed (Case and Swanson 2002). Nevertheless, MTF have the potential to cover a broader range of content per question than Type A (Dudley 2006). As this enables more information per testing time to be revealed, higher test reliabilities have been found when using MTF compared to Type A (Frisbie and Sweeney 1982; Javid 2014; Kreiter and Frisbie 1989; Mobalegh and Barati 2012; Siddiqui et al. 2016). Moreover, in contrast to a common prejudice, MTF items are capable of measuring higher cognitive levels, such as comprehension according to Bloom's taxonomy (Downing and Yudkowsky 2009; Richardson 1992).

Overall, MTF items show advantages over other multiple response formats (like Pick-N and MM), such as minimizing test-wiseness, and therefore seem to be a valuable supplement to Type A items. They are commonly used in medical school exams as well as licensing examinations such as the Swiss Federal Licensing Exam (FLE) (Guttormsen et al. 2013) or the US National Certification and Licensure Exam for Registered Nurses (NCLEX-RN) (Dunham 2006).

The validity of a test is influenced, among other things, by the scoring of the items (Cook et al. 2015). When using MTF items for high-stakes assessment, evidence about the optimal scoring algorithm is necessary. Different scoring algorithms exist for scoring MTF items, which mainly differ in whether partial knowledge is rewarded and whether there is a penalty for wrong answers. In "dichotomous scoring" (*DS*), no partial knowledge is rewarded and examinees only receive a point if all answer options of an item are marked correctly. *DS* is the strictest scoring algorithm, following the rationale that patient treatment cannot tolerate mistakes and that especially qualifying exams have to test accordingly (Albanese and Sabers 1988; Bauer et al. 2011; Itten and Krebs 1997; Verbić 2012). In

partial credit scoring (PS), partial knowledge is rewarded. Different forms of PS exist: In the most basic form, examinees receive an equal fraction of a point for every correctly marked answer option (Albanese and Sabers 1988; Itten and Krebs 1997; Verbić 2012). This scoring algorithm follows the understanding that every correct choice might be beneficial for the patient (Bauer et al. 2011). PS can also take into consideration different thresholds of knowledge, grounded on the assumption that a mere fraction of knowledge is not sufficient to treat a patient adequately. While a certain portion of knowledge allows the examinee to make decisions that are not entirely wrong, he or she might be missing out on better alternatives (Bauer et al. 2011). Sometimes, *negative marking* is used, which penalizes examinees for wrong answers in order to prevent guessing (Siddiqui et al. 2016). However, this scoring algorithm lowers validity, as it introduces construct-irrelevant factors like personality and response styles (Gross 1982).

Studies examining the psychometric characteristics of differently scored MTF have yielded contradictory findings. While three studies found that partial credit scoring shows better psychometric characteristics, such as higher reliabilities, than dichotomous scoring (Albanese and Sabers 1988; Itten and Krebs 1997; Krebs 1997), two studies found no differences in the psychometric characteristics of different scoring algorithms (Tsai and Suen 1993; Wu 2003). Four of the above-mentioned studies were conducted using summative assessments such as end-of-term assessment (Albanese and Sabers 1988; Itten and Krebs 1997; Krebs 1997) or college entrance assessments (Wu 2003), while the remaining study did not report whether the analyzed assessment had an impact on the students' study progress (Tsai and Suen 1993).

As a robust scoring algorithm is essential for the valid interpretation of exam data, and the existing evidence is scarce, we aim to clarify the optimal scoring of MTF using a large number of exams. Our research question is as follows: What is the optimal scoring algorithm for MTF items regarding reliability, difficulty index and item-total correlation?

As Type A items are one of the most widely used item formats in achievement testing, it is of practical importance to document the effectiveness of any other item format in relation to Type A (Kreiter and Frisbie 1989). Therefore, we also investigate the psychometric properties of Type A items used in the same exams in our study to enable us to compare our results for MTF with this widely used item format.

Methods

To clarify the influence of scoring on psychometric characteristics of exams, we analyzed empirical data from 37 exams covering the range from first-year end-of-term exams to the Swiss Federal Licensing Exam (FLE) after the sixth year. Exams were conducted at two Swiss medical schools in 2015. The FLE is organized centrally and conducted at five Swiss medical schools.

The mean number of examinees per exam was 243 (SD = 82; Min = 150; Max = 887). Exams consisted of both MTF items (M = 26.97; SD = 6.1; Min = 15; Max = 43; total = 998) and Type A (M = 58.46; SD = 40.5; Min = 23; Max = 270; total = 2163). Each Type A item included five answer options and each MTF item contained four answer options. Items eliminated in a post hoc review were excluded from the analyses (2.19 items per exam on average). To compare item characteristics, MTF subsets as well as Type A subsets were treated as stand-alone tests. All blueprint categories were covered with MTF and Type A alike, improving the validity of subsets and allowing for a

better comparison of Type A and MTF items. We compared three item scoring algorithms for MTF items (cf. Table 1):

- Dichotomous Scoring (DS): Examinees receive a full point only if the true/false rating of all four options was marked correctly.
- Partial Scoring (PS_{50}): Examinees receive half a point if more than half of true/false ratings were marked correctly and one point if all were marked correctly.
- Partial Scoring ($PS_{1/n}$): Examinees receive a quarter of a point for every correct true/false rating.

As Type A questions have only one correct solution, they were scored with a full point when answered correctly; otherwise, candidates received no points.

The MTF items included in this study were originally scored with PS_{50}. We simulated the other two scoring algorithms over the existing data. We did not simulate negative marking in our study, as we assumed that examinees might have shown a different response behavior.

Reliability was operationalized as Cronbach's alpha calculated for each combination of MTF subset and scoring algorithm as well as for the Type A subsets. These empirical data were then projected to 50 item test sets using the Spearman-Brown prediction formula to adjust for different lengths of the exams. To stabilize Cronbach's alpha coefficients for analyses, we used Fisher's z-transformation (Romano et al. 2010). Using the Spearman-Brown prediction formula, we additionally calculated the number of items needed for a reliability of 0.8, considered a minimal requirement for end of term exams (Downing and Yudkowsky 2009).

Further psychometric criteria of interest were mean *item difficulty* and *item discrimination*. *Item difficulty* describes the average score of an item. In dichotomously scored items, this is equal to the proportion of examinees who answer an item correctly. The interpretation of the item difficulty is somewhat counterintuitive, as higher item difficulty indicates easier items. Items with medium difficulty are most informative. According to the recommendations of Downing and Yudkowsky (2009) we defined for our analysis items under 0.24 as extremely difficult and items over 0.91 as extremely easy. *Item discrimination* describes how well an item discriminates between higher-performing and lower-performing test takers. Thereby high positive discrimination is always better than low negative discrimination. For this analysis, we defined discrimination indices over 0.2 as sufficient (Downing and Yudkowsky 2009). Item discrimination was operationalized as mean Pearson product-moment correlation coefficient (for further information: Downing and Yudkowsky 2009).

Psychometric indicators of the differently scored tests as well as those of Type A were compared using repeated measures analyses of variance (rANOVA) with Bonferroni correction for multiple comparisons. Thereby, we used exams, not single items as a data

Table 1 Received points for different numbers of correct answers for different scoring algorithms for MTF items with four answer options

Number of correct choices	0	1	2	3	4
DS	0.00	0.00	0.00	0.00	1.00
PS_{50}	0.00	0.00	0.00	0.50	1.00
$PS_{1/n}$	0.00	0.25	0.50	0.75	1.00

basis for the rANOVA. As control variables, we included the medical schools and level of training. Studies indicate differences in gender when using negative marking (Baldiga 2013; Ravesloot et al. 2015). As we did not include negative marking in our study, we did not expect differences in gender, and therefore did not include gender as a control variable. Additionally, we calculated Spearman's rank correlation to control the rank order. Correlation coefficients were corrected for unreliability (Muchinsky 1996). All analyses were conducted using R (version 3.2.0) (R Core Team 2013). As index for effect sizes, we calculated partial η^2.

The ethics committee of the Canton of Bern (Switzerland) declared that no vote was necessary for this educational study. We confirm that the original test takers cannot be identified by the material presented and that they undergo no conceivable risk by having their test data included in this study. The medical schools and the FLE examination board gave their permission to use the data.

Results

Test reliability

When projected to 50 item exams, test reliability ranged from 0.70 for DS to 0.75 for PS_{50} and $PS_{1/n}$. Reliability of PS_{50} and $PS_{1/n}$ was significantly higher compared to DS and (dichotomously scored) Type A [$F(3,108) = 22.72$; sig < 0.000; $\eta^2 = 0.034$]. Differences between PS_{50} and $PS_{1/n}$ as well as differences between DS and Type A were not significant.

To arrive at a Cronbach's alpha of 0.8, significantly fewer items are needed with PS_{50} and $PS_{1/n}$ than with DS and Type A [$F(3,108) = 14.37$; sig < 0.000; $\eta^2 = 0.285$]. Differences between PS_{50} and $PS_{1/n}$ as well as differences between DS and Type A were not significant. Detailed results are shown in Table 2.

Item difficulty

Item difficulties differed significantly depending on the scoring algorithm [$F(3,108) = 1076.39$; sig < 0.000; $\eta^2 = 0.791$]. The mean difficulty ranged from 0.5 for DS to 0.82 for $PS_{1/n}$, demonstrating that items with $PS_{1/n}$ were easier. The difficulty of Type A items lay between that of PS_{50} and $PS_{1/n}$.

When analyzing the number of items that were too easy ($p > .91$) or too difficult ($p < .24$), we found significant differences in all scoring algorithms

Table 2 Mean reliabilities projected to 50 items (α_{50}), mean z-value for reliability and number of items needed for a Cronbach's alpha of 0.8 ($n_{0.8}$)

Scoring algorithm	α_{50}	SD (α_{50})	z	SD (z)	$n_{0.8}$	SD ($n_{0.8}$)
DS	0.70	0.15	0.93	0.32	103	78.8
PS_{50}	0.75	0.13	1.05	0.33	75	54.1
$PS_{1/n}$	0.75	0.13	1.06	0.32	74	52.7
Type A	0.72	0.12	0.96	0.30	87	51.7

[$F_{p>0.9}(3,108) = 45.92$; $sig_{p>0.9} < 0.000$; $\eta^2_{p>0.9} = 0.561$ $F_{p<0.24}(3,108) = 45{,}64$; $sig_{p<0.24} < 0.000$ $\eta^2_{p>0.24} = 0.559$]. The use of DS led to the highest number of items that were too difficult, while $PS_{1/n}$ led to the highest number of items that were too easy. Detailed results are shown in Table 3.

Item discrimination

The use of $PS_{1/n}$ and PS_{50} resulted in significantly higher item-total correlations compared to DS and Type A [$F(3,108) = 60.21$; sig < 0.000; $\eta^2 = 0.626$]. Differences between $PS_{1/n}$ and PS_{50} as well as differences between DS and Type A were not significant. Item discrimination was sufficient in all scoring algorithms. Detailed results are shown in Table 4.

Influencing variables

The control variables (medical school, year of study) showed no significant interaction in our analyses, indicating that none of them had an impact on the results [medical school: reliability: $F(6,102) = 1.12$; sig $= 0.35$; difficulty: $F(6,102) = 0.22$; sig $= 0.65$; discrimination: $F(6,102) = 0.44$; sig $= 0.59$; year of study: reliability: $F(3,105) = 0.61$; sig $= 0.28$; difficulty: $F(3,105) = 0.29$; sig $= 0.72$; discrimination: $F(3,105) = 1.37$; sig $= 0.26$].

Covariates

Students' rank order in the different scoring algorithms correlated highly with each other. This became even more salient when correcting the coefficients for unreliability, and indicates that examinees' rank order did not change when changing the scoring algorithm. Additionally, we found high correlation coefficients between students' rank order in the different MTF scoring algorithms and in Type A items. Detailed results are shown in Table 5.

Discussion

Previous studies reported conflicting evidence regarding the psychometric properties of differently scored MTF items. We included 37 exams, to clarify the effects of different scoring algorithms for MTF items on psychometric properties and compared the

Table 3 Mean item difficulty-index and number of extremely difficult and extremely easy items for each scoring algorithm

Scoring algorithm	Mean (p)	SD (p)	$p > 0.91$ (%)	$p < 0.24$ (%)
DS	0.50	0.06	20 (2.0)	136 (13.6)
PS_{50}	0.66	0.07	61 (6.1)	22 (2.2)
$PS_{1/n}$	0.82	0.08	196 (19.6)	2 (0.2)
Type A	0.73	0.08	387 (17.9)	45 (2.1)

Table 4 Mean discrimination-index and number of *not* discriminating items for each scoring algorithm

Scoring algorithm	Mean (r)	SD (r)	$r < 0.2$
DS	0.30	0.06	0
PS_{50}	0.33	0.07	0
$PS_{1/n}$	0.33	0.08	0
Type A	0.28	0.08	2

Table 5 Rank correlations of candidates' points in the different scoring algorithms and in Type A

	PS_{50}	$PS_{1/n}$	Type A
DS	0.975** (1.16[a])	0.963** (1.15[a])	0.776** (.940[a])
PS_{50}		0.996** (1.16[a])	0.823** (.975[a])
$PS_{1/n}$			0.853** (1.00[a])

[a]Correlations were corrected for unreliability

**$p < .001$

psychometric results with those of Type A questions. Results were robust and were influenced neither by the medical school nor by the level of training.

Candidates' points received with different scoring algorithms for MTF items were almost perfectly correlated. This indicates that MTF items measure the same ability independently of the scoring algorithm used. As the scoring algorithms were simulated over the same items, this result is to be expected. All blueprint categories were covered with both MTF and Type A. Correlations between MTF items and Type A were high, indicating that the two types of MC questions measure similar abilities.

In general, partial scoring (PS) algorithms showed higher discrimination indices and slightly higher reliability, implicating that less items are needed for a reliability of 0.8 than dichotomous scoring (DS) and Type A. The higher reliability for partial scoring is in line with findings from the literature. Partial scoring takes partial knowledge into account, which can lead to a more precise discrimination of examinees, while partial knowledge remains invisible in dichotomously scored MTF and in Type A items. On the other side, partial scoring is also more prone to acknowledge chance success compared to dichotomously scored MTF and Type A items. The increased reliability for partial scoring algorithms would support the argument that partial information contributes more to measurement precision than does reducing chance success. Comparing the two partial scoring algorithms, we found no differences in reliability and discrimination. We hypothesize that the gain in partial information obtained if a candidate answers fewer than three of the four options of an MTF item correctly is offset by the increased credit of chance success.

We found that PS_{50} showed good difficulties and covered the whole range from easy to difficult items. This indicates that we included carefully constructed items in our study. The difficulty of the Type A subsets was consistently close to the PS_{50}-scored MTF items.

When changing the scoring algorithm from PS_{50} to $PS_{1/n}$, items became easier, whereas applying DS led to more difficult items. Items that are too easy or too difficult do not

differentiate between examinees. Unless their content is essential, they tend to be a waste of resources (Downing and Yudkowsky 2009) and should therefore be avoided. To that end, PS_{50} may have advantages, as it covers the optimal range without producing too many difficult or easy items. However, this finding might be biased, as items were originally constructed for PS_{50}, and the other two scoring algorithms were only retrospectively simulated over the items. We chose this approach of simulating DS and $PS_{1/n}$ over the existing data because it enabled us to use "real exams that count" (Norman et al. 1996). However, having a scoring algorithm in mind might influence authors' item development, and consequently the psychometric characteristics. In any case, authors should be informed about which scoring algorithm will be used and existing items would have to be adapted to changes in scoring. As items become easier, for example, with $PS_{1/n}$, authors may tend to test uncommon specialist knowledge in order to produce more difficult items. Further research could analyze whether and how authors are mindful of the intended scoring when constructing items.

In a nutshell, partial credit scoring leads to better psychometric results than dichotomous scoring and produces similar results to Type A. Until it has been demonstrated that authors are able to adapt item development to the intended scoring algorithm, so that an exam with $PS_{1/n}$ scored items also covers the whole range of item difficulty, we suggest using PS_{50}.

Implications for practice and research

Analyzing the effect of different scoring algorithms for multiple true–false (MTF) items on psychometric characteristics, partial credit scoring showed slightly higher reliability and discrimination than dichotomous scoring. Even though differences in reliability are small, a notable implication for practice is that considerable fewer items are needed to achieve a reliability of 0.8 when using partial credit scoring. Regarding item difficulty, PS_{50} showed the best results, covering the whole range from low to high difficulty. In conclusion, we would recommend partial credit scoring as a scoring algorithm for MTF items.

Acknowledgements The authors thank the examination board of the Swiss Federal Licensing Examination as well as the two Swiss medical schools for providing the data from the included exams. The authors wish to express their gratitude to the editor for the helpful guidance during the review process.

References

Albanese, M. A., & Sabers, D. L. (1988). Multiple true–false items: A study of interitem correlations, scoring alternatives, and reliability estimation. *Journal of Educational Measurement, 25*(2), 111–123.

Baldiga, K. (2013). Gender differences in willingness to guess. *Management Science, 60*(2), 434–448.

Bauer, D., Holzer, M., Kopp, V., & Fischer, M. R. (2011). Pick-N multiple choice-exams: A comparison of scoring algorithms. *Advances in Health Sciences Education, 16*(2), 211–221.

Case, S. M., & Swanson, D. B. (2002). *Constructing written test questions for the basic and clinical sciences* (3rd ed.). Philadelphia, PA: National Board of Medical Examiners.

Cook, D. A., Brydges, R., Ginsburg, S., & Hatala, R. (2015). A contemporary approach to validity arguments: A practical guide to Kane's framework. *Medical Education, 49*(6), 560–575. https://doi.org/10.1111/medu.12678.

Cronbach, L. (1939). Note on the multiple true–false test exercise. *Journal of Educational Psychology, 30*(8), 628.

Cronbach, L. (1941). An experimental comparison of the multiple true–false and multiple multiple-choice tests. *Journal of Educational Psychology, 32*(7), 533.

Downing, S. M., & Yudkowsky, R. (2009). *Assessment in health professions education*. New York: Routledge.

Dudley, A. (2006). Multiple dichotomous-scored items in second language testing: Investigating the multiple true–false item type under norm-referenced conditions. *Language Testing, 23*(2), 198–228.

Dunham, M. L. (2006). *An investigation of the multiple true–false item for nursing licensure and potential sources of construct-irrelevant difficulty*. ProQuest.

Frisbie, D. A., & Sweeney, D. C. (1982). The relative merits of multiple true–false achievement tests. *Journal of Educational Measurement, 19*(1), 29–35. https://doi.org/10.2307/1434916.

Gross, L. J. (1982). Scoring multiple true/false tests some considerations. *Evaluation and the Health Professions, 5*(4), 459–468.

Guttormsen, S., Beyeler, C., Bonvin, R., Feller, S., Schirlo, C., Schnabel, K., et al. (2013). The new licencing examination for human medicine: From concept to implementation. *Swiss Medical Weekly, 143,* w13897. https://doi.org/10.4414/smw.2013.13897.

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15*(3), 309–333.

Itten, S., & Krebs, R. (1997). *Messqualität der verschiedenen MC-Itemtypen in den beiden Vorprüfungen des Medizinstudiums an der Universität Bern 1997/2 (Forschungsbericht Institut für Aus-, Weiter-und Fortbildung (IAWF) der medizinischen Fakultät der Universität Bern).* Bern: IAWF.

Javid, L. (2014). The comparison between multiple-choice (MC) and multiple true–false (MTF) test formats in iranian intermediate EFL learners' vocabulary learning. *Procedia-Social and Behavioral Sciences, 98,* 784–788.

Krebs, R. (1997). The swiss way to score multiple true–false items: theoretical and empirical evidence. In A. J. J. A. Scherpbier, C. P. M. van der Vleuten, J. J. Rethans, & A. F. W. van der Steeg (Eds.), *Advances in medical education* (pp. 158–161). Netherlands: Springer.

Krebs, R. (2004). *Anleitung zur Herstellung von MC-Fragen und MC-Prüfungen für die ärztliche Ausbildung*. Bern: Institut für Medizinische Lehre IML, Abteilung für Ausbildungs-und Examensforschung AAE.

Kreiter, C. D., & Frisbie, D. A. (1989). Effectiveness of multiple true–false items. *Applied Measurement in Education, 2*(3), 207–216.

Mobalegh, A., & Barati, H. (2012). Multiple true–false (MTF) and multiple-choice (MC) test formats: A comparison between two versions of the same test paper of Iranian NUEE. *Journal of Language Teaching and Research, 3*(5), 1027–1037.

Muchinsky, P. M. (1996). The correction for attenuation. *Educational and Psychological Measurement, 56*(1), 63–75.

Norman, G. R., Swanson, D. B., & Case, S. M. (1996). Conceptual and methodological issues in studies comparing assessment formats. *Teaching and Learning in Medicine: An International Journal, 8*(4), 208–216.

R Core Team. (2013). R: A language and environment for statistical computing. Vienna, Austria. Retrieved from http://www.r-project.org/.

Ravesloot, C., Van der Schaaf, M., Muijtjens, A., Haaring, C., Kruitwagen, C., Beek, F., et al. (2015). The don't know option in progress testing. *Advances in Health Sciences Education, 20*(5), 1325–1338.

Richardson, R. (1992). The multiple choice true/false question: What does it measure and what could it measure? *Medical Teacher, 14*(2–3), 201–204.

Romano, J. L., Kromrey, J. D., & Hibbard, S. T. (2010). A Monte Carlo study of eight confidence interval methods for coefficient alpha. *Educational and Psychological Measurement, 70,* 376–393.

Siddiqui, N. I., Bhavsar, V. H., Bhavsar, A. V., & Bose, S. (2016). Contemplation on marking scheme for Type X multiple choice questions, and an illustration of a practically applicable scheme. *Indian Journal of Pharmacology, 48*(2), 114.

Tarasowa, D., & Auer, S. (2013). *Balanced scoring method for multiple-mark questions.* Paper presented at the CSEDU.

Tsai, F.-J., & Suen, H. K. (1993). A brief report on a comparison of six scoring methods for multiple true–false items. *Educational and Psychological Measurement, 53*(2), 399–404.

Verbić, S. (2012). Information value of multiple response questions. *Psihologija, 45*(4), 467–485.

Wu, B. C. (2003). Scoring multiple true false items: A comparison of summed scores and response pattern scores at item and test levels. Retrieved from Eric: https://eric.ed.gov/?id=ED476148