Scoring Single-Response Multiple-Choice Items – Quite Simple?! A Scoping Review and Comparison of Different Scoring Methods

Amelie Friederike Kanzow, Dennis Schmidt, Philipp Kanzow

Submitted to: JMIR Medical Education
on: November 05, 2022

Table of Contents

Scoring Single-Response Multiple-Choice Items – Quite Simple?! A Scoping Review and Comparison of Different Scoring Methods

Amelie Friederike Kanzow[1] MEd; Dennis Schmidt[2] MSc; Philipp Kanzow[2] MSc, Dr rer medic, PD Dr med dent

[1]Study Deanery University Medical Center Göttingen Göttingen DE
[2]Department of Preventive Dentistry, Periodontology and Cariology University Medical Center Göttingen Göttingen DE

Corresponding Author:
Philipp Kanzow MSc, Dr rer medic, PD Dr med dent
Department of Preventive Dentistry, Periodontology and Cariology
University Medical Center Göttingen
Robert-Koch-Str 40
Göttingen
DE

Abstract

Background: Single-choice items (eg, <i>best-answer items</i>, <i>alternate-choice items</i>, <i>single true-false items</i>) are one type of multiple-choice items and have been used in examinations for over 100 years. At the end of every examination, the examinees' responses have to be analyzed and scored in order to derive with an information about examinees' <i>true knowledge</i>.

Objective: The aim of this paper is to compile scoring methods for individual single-choice items described in the literature. Furthermore, the metric <i>expected chance score</i> and the relation between examinees' <i>true knowledge</i> and expected scoring results (averaged percentage score) are analyzed. Furthermore, implications for potential pass marks to be used in examinations to test examinees for a predefined level of <i>true knowledge</i> are derived.

Methods: Scoring methods for individual single-choice items including were extracted from various databases (ERIC, PsycInfo, Embase via Ovid, MEDLINE via PubMed) in September 2020. Eligible sources reported on scoring methods for individual single-choice items in written examinations including but not limited to medical education. Separately for items with <i>n</i> = 2 answer options (eg, <i>alternate-choice items</i>, <i>single true-false items</i>) and <i>best-answer items</i> with <i>n</i> = 5 answer options (eg, <i>Type A</i> items) and for each identified scoring method, the metric <i>expected chance score</i> and the expected scoring results as a function of examinees' <i>true knowledge</i> using fictitious examinations with 100 single-choice items were calculated.

Results: A total of 21 different scoring methods were identified from the 258 included sources, with varying consideration of correctly marked, omitted, and incorrectly marked items. Resulting credit varied between -3 and +1 credit points per item. For items with <i>n</i> = 2 answer options, <i>expected chance scores</i> from random guessing ranged between -1 and +0.75 credit points. For items with <i>n</i> = 5 answer options, <i>expected chance scores</i> ranged between -2.2 and +0.84 credit points. All scoring methods showed a linear relation between examinees' <i>true knowledge</i> and the expected scoring results. Depending on the scoring method used, examination results differed considerably: Expected scoring results from examinees with 50% <i>true knowledge</i> ranged between 0.0% (95% CI: 0% to 0%) and 87.5% (95% CI: 81.0% to 94.0%) for items with <i>n</i> = 2 and between -60.0% (95% CI: -60% to -60%) and 92.0% (95% CI: 86.7% to 97.3%) for items with <i>n</i> = 5.

Conclusions: In examinations with single-choice items, the scoring result is not always equivalent to examinees' <i>true knowledge</i>. When interpreting examination scores and setting pass marks, the number of answer options per item must usually be taken into account in addition to the scoring method used.

(JMIR Preprints 05/11/2022:44084)
DOI: https://doi.org/10.2196/preprints.44084

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

Please make my preprint PDF available to anyone at any time (recommended).

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.

Only make the preprint title and abstract visible.

✓ **No, I do not wish to publish my submitted manuscript as a preprint.**

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Review

Scoring Single-Response Multiple-Choice Items – Quite Simple?! A Scoping Review and Comparison of Different Scoring Methods

Abstract

Background: Single-choice items (eg, *best-answer items*, *alternate-choice items*, *single true-false items*) are one type of multiple-choice items and have been used in examinations for over 100 years. At the end of every examination, the examinees' responses have to be analyzed and scored in order to derive with an information about examinees' *true knowledge*.

Objectives: The aim of this paper is to compile scoring methods for individual single-choice items described in the literature. Furthermore, the metric *expected chance score* and the relation between examinees' *true knowledge* and expected scoring results (averaged percentage score) are analyzed. Furthermore, implications for potential pass marks to be used in examinations to test examinees for a predefined level of *true knowledge* are derived.

Methods: Scoring methods for individual single-choice items including were extracted from various databases (ERIC, PsycInfo, Embase via Ovid, MEDLINE via PubMed) in September 2020. Eligible sources reported on scoring methods for individual single-choice items in written examinations including but not limited to medical education. Separately for items with n = 2 answer options (eg, *alternate-choice items*, *single true-false items*) and *best-answer items* with n = 5 answer options (eg, *Type A* items) and for each identified scoring method, the metric *expected chance score* and the expected scoring results as a function of examinees' *true knowledge* using fictitious examinations with 100 single-choice items were calculated.

Results: A total of 21 different scoring methods were identified from the 258 included sources, with varying consideration of correctly marked, omitted, and incorrectly marked items. Resulting credit varied between -3 and +1 credit points per item. For items with *n* = 2 answer options, *expected chance scores* from random guessing ranged between -1 and +0.75 credit points. For items with *n* = 5 answer options, *expected chance scores* ranged between -2.2 and +0.84 credit points. All scoring methods showed a linear relation between examinees' *true knowledge* and the expected scoring results. Depending on the scoring method used, examination results differed considerably: Expected scoring results from examinees with 50% *true knowledge* ranged between 0.0% (95% CI: 0% to 0%) and 87.5% (95% CI: 81.0% to 94.0%) for items with *n* = 2 and between -60.0% (95% CI: -60% to -60%) and 92.0% (95% CI: 86.7% to 97.3%) for items with *n* = 5.

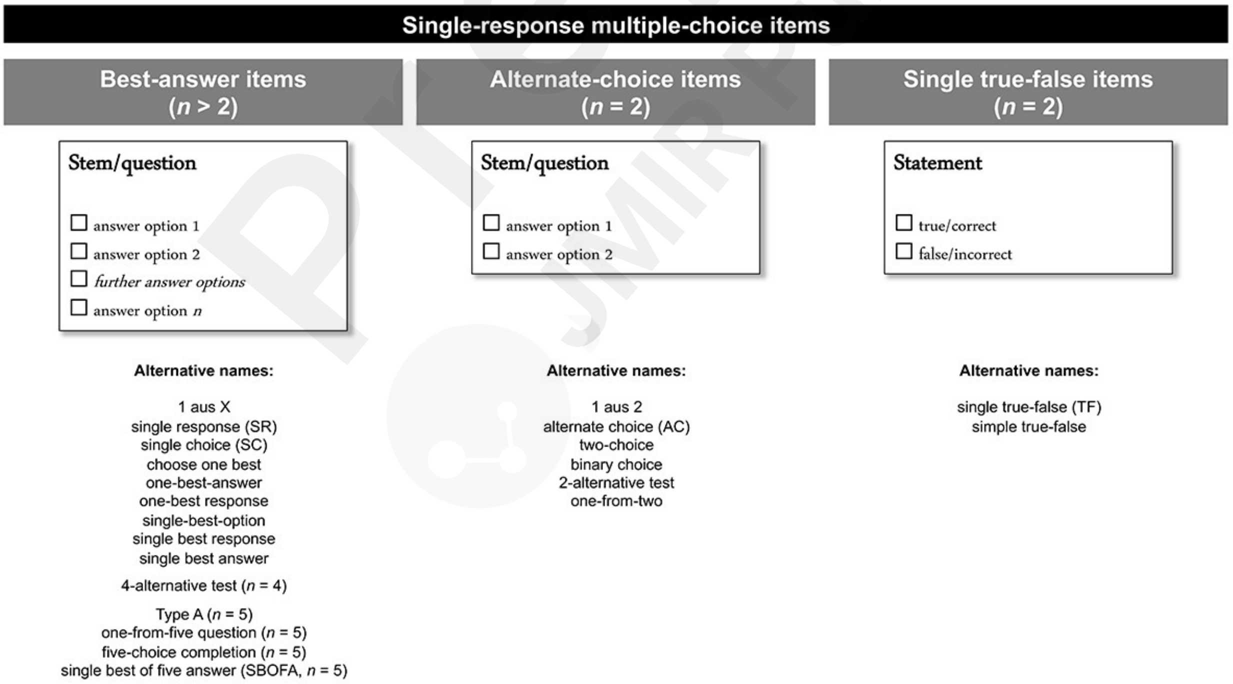
Conclusions: In examinations with single-choice items, the scoring result is not always equivalent to examinees' *true knowledge*. When interpreting examination scores and setting pass marks, the number of answer options per item must usually be taken into account in addition to the scoring method used.

Keywords: alternate-choice; best-answer; education; education system; educational assessment; educational measurement; examination; multiple-choice; results; scoring; scoring system; single-choice; single-response; scoping review; test; testing; true/false; true-false; type A

Introduction

Multiple-choice items in single-response item formats (ie, *single-choice items*) require examinees to mark only one answer option or to make only one decision per item. The most frequently used item type among the group of single-choice items are so-called *best-answer items*. Here, examinees must select exactly one (ie, the correct or most likely) answer option from the given answer options [1]. Often, *best-answer items* contain five answer options ($n = 5$), although the number of answer options might vary ($n \geq 2$). Items with exactly two answer options ($n = 2$) are also referred to as alternative items (ie, *alternate-choice items*) [2]. In addition, *single true-false items* belong to the group of single-choice items. Examples of the mentioned single-choice items as well as alternative designations are shown in Figure 1.

Figure 1. Examples of three different multiple-choice items in single-choice format and alternative designations used in the literature (no claim to completeness).



Single-choice items have been used for more than 100 years to test examinees' knowledge. The use of these items began among U. S. school pupils, which were given *alternate-choice* or *best-answer items* [3] or *single true-false items* [4] as a time-saving alternative to conventional open-ended questions (ie, *essay type examinations*). Due to their character of only allowing clearly correct or incorrect responses from examinees, multiple-choice examinations were also called *objective type examinations* [5]. The term *new type examinations* was coined to distinguish them from previously commonly used open-ended questions [5, 6].

The use of multiple-choice items did not remain exclusive to the setting of high schools but also extended to examinations in university contexts [7] and postgraduate medical education [8, 9]. Today, multiple-choice items are frequently used in examinations of medical and dental students (eg, within the *United States Medical Licensing Examination*). Besides their usage in individual medical or dental programmes, different multiple-choice item types found their way into examinations for medical students by the *National Board of Medical Examiners* [10]: Within the context of single-choice items, single-choice items with n = 5 were particularly used and referred to as *Type A* items.

Examinations aim at assessing examinees' ability (ie, examinees' *true knowledge* [k]) regarding predefined learning objectives. The downside when using multiple-choice examinations is that examinees might also mark an item correctly by guessing or by identifying the correct answer option through recognition. Thus, an active knowledge reproduction does not necessarily take place, and correct responses are not necessarily resulting from examinees' *true knowledge*.

To grade examinees or to decide about passing or failing a summative examination based on a minimum required level of *true knowledge*, scoring algorithms are used to transfer examinees' responses (ie, marking schemes) into a score. To assess examinees' *true knowledge*, the obtained scores must either be reduced by the guessing factor, negative points (ie, malus points) must be assigned for incorrectly marked items, and/or the pass mark (ie, the corresponding cut-off score for the desired *true knowledge* cut-off value) must be adjusted based on the guessing probability [11]. The guessing probability for examinees without any knowledge (k = 0, blind guessing only) amounts to 20% for *single-choice items* with n = 5 and to 50% for *alternate-choice items* and *single true-false items* with n = 2. Consequently, examinees without any knowledge score 20% or 50% of the maximum score on average, respectively [11]. However, it can be assumed that most examinees have at least partial knowledge ($0 < k < 1$) and that an informed guessing with remaining partial uncertainty occurs in most cases.

Since the introduction of multiple-choice items, numerous scoring methods have been described in the literature and (medical) educators are advised to choose an appropriate scoring method based on an informed decision. Therefore, the aim of this scoping review is (1) to map an overview of different scoring methods for individual single-choice items described in the literature, (2) to compare different scoring methods based on the metric *expected chance score*, and (3) to analyze the relation between examinees' *true knowledge* and expected scoring results (averaged percentage score).

Methods

Systematic Literature Search

The literature search was performed according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews Checklist [12]. The checklist is available as Multimedia Appendix 1. As this review did not focus on health outcomes, the review was not registered at PROSPERO prior to its initiation.

Eligibility criteria

Potentially eligible sources included scientific articles, books, book chapters, dissertations, and congress abstracts reporting scoring methods for individual single-choice items in written examinations including but not limited to medical examinations. Scoring methods for item groups or scoring on examination level (eg, with different weighting of individual items, with mixed item types or considering the total number of items per examination) were not assessed. Also, scoring methods that deviate from the usual marking procedure (ie, a single choice of marking exactly one answer option per item) were not considered. These include, for example, procedures that assess the confidence of examinees in their marking (eg, *confidence weighting*), let examinees select the incorrect answer options (eg, *elimination scoring*), let examinees narrow down the correct answer option (eg, *subset selection*), or allow for the correction of initially incorrectly marked items (eg, *answer-until-correct*). No further specifications were made regarding language, quality (eg, minimum impact factor), or time of publication.

Information Sources

Four databases (ERIC, PsycInfo, Embase via Ovid, and MEDLINE via PubMed) were searched in September 2020. The search term was composed of various designations for single-choice items as well as keywords with regard to examinations. It was slightly adapted according to the specifications of the individual databases. The respective search terms for each database can be found in Table 1.

Table 1. Search terms used for each of the four databases.

Database	Search Term
ERIC	("single choice" OR "alternate choice" OR "single response" OR "one-best-answer" OR "single best response" OR "true-false" OR "Typ A")

	AND (item OR items OR test OR tests OR testing OR score OR scoring OR examination OR examinations)
PsycInfo	("single choice" OR "alternate choice" OR "single response" OR "one-best-answer" OR "single best response" OR "true-false" OR "Typ A") AND (item OR items OR test OR tests OR testing OR score OR scoring OR examination OR examinations)
Embase via Ovid	(("single choice" or "alternate choice" or "single response" or "one-best-answer" or "single best response" or "true-false" or "Typ A") and (item OR items or test or tests or testing or score or scoring or examination or examinations)).af.
MEDLINE via PubMed	("single choice"[All Fields] OR "alternate choice"[All Fields] OR "single response"[All Fields] OR "one-best-answer" OR "single best response" OR "true-false"[All Fields] OR "Typ A"[All Fields]) AND ("item"[All Fields] OR "items"[All Fields] OR "test"[All Fields] OR "tests"[All Fields] OR "testing"[All Fields] OR "score"[All Fields] OR "scoring"[All Fields] OR "examination"[All Fields] OR "examinations"[All Fields])

Selection of Sources

Literature screening, inclusion of sources, and data extraction was independently performed by two authors (AFB and PK). First, the titles and abstracts of the database results were screened. Duplicate results as well as records being irrelevant to the research question were sorted out. For books and book chapters, however, different editions were included separately. In a second step, full-texts sources were screened, and eligible records were included as sources. In addition, the references of included sources were searched in an additional hand search for further, potentially relevant sources. After each step, the results were compared, and any discrepancies were discussed until a consensus was reached. Information with regard to the described scoring methods were extracted using a piloted checklist.

Data Extraction

The following data were extracted from included sources using a piloted spreadsheet if reported: (1) name of the scoring method, (2) associated item type, and (3) algorithm for calculating scores per item. The mathematical equations of each scoring method were adjusted to achieve normalization of scores up to a maximum of +1 point per item if necessary.

Data Synthesis

For all identified scoring methods, the expected scoring results in case of pure guessing were calculated for single-choice items with n = 2 and n = 5 answer options, respectively [13]. The *expected chance score* is described in the literature as a comparative metric of different scoring methods [11, 13-15]. For its calculation, examinees without any knowledge (k = 0) are expected to always guess blindly and thus achieve the *expected chance score* on average.

In addition, expected scoring results for varying levels of k ($0 \leq k \leq 1$) were calculated. For examinees with partial knowledge ($0 < k < 1$), a correct response can be attributed to both partial knowledge and guessing, with the proportion of guessing decreasing as knowledge increases. In contrast, examinees with perfect knowledge (k = 1) always select the correct answer option without the need for guessing [11].

Examinees were expected to answer all items, and it was supposed that examinees were unable to omit individual items or that examinees do not use an omit option. Furthermore, all items and answer options were assumed to be of equal difficulty and to not contain any cues. The equation for the calculation of the expected scoring result is shown in Figure 3.

Figure 3. Equation for the calculation of the expected scoring result (f = credit points awarded for a correctly marked item [i = 1] or an incorrectly marked item [i = 0] depending

on the scoring method used; k = examinees' *true knowledge* [$0 \leq k \leq 1$]; n = number of answer options per item; $x = 1$ if the correct answer option is selected by *true knowledge*, otherwise $x = 0$; in the equation shown, 0^0 is defined as 1).

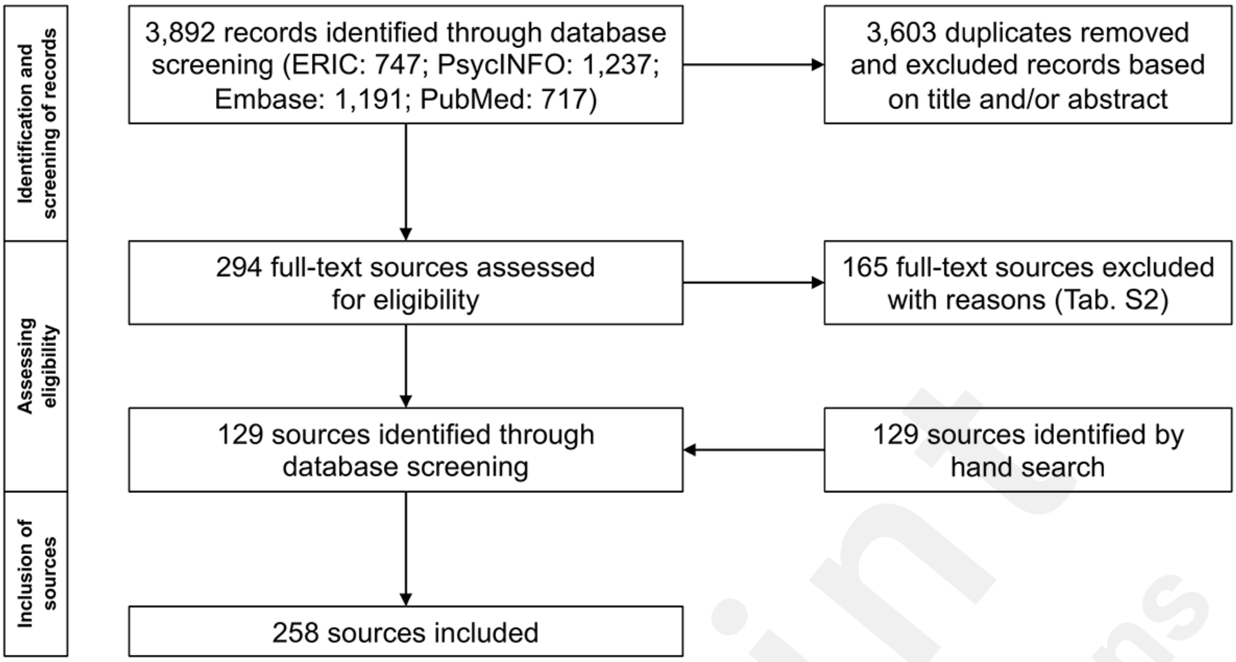
$$\mathit{Expected\ scoring\ result} = \sum_{i=0}^{1}\sum_{x=0}^{i}\left(k^{x} * (1-k)^{1-x}\right) * \frac{\binom{n-1}{1-i}}{\binom{n-x}{1-x}} * f_i$$

MATLAB software (version R2019b; The MathWorks, Natick, MA, USA) was used to calculate the relation between examinees' *true knowledge* and the expected scoring results using fictitious examinations consisting of 100 single-choice items (all items with either n = 2 or n = 5).

Results

Within the literature search, a total of 3,892 records were found through database search. Of these, 129 sources could be included. A further 129 sources were identified from the references of the included sources by hand search. The entire process of screening and including sources is shown in Figure 2. Reasons for exclusion of sources during full-text screening are given in Multimedia Appendix 2.

Figure 2. Flow diagram of systematic literature search.


Identification and screening of records
Assessing eligibility
Inclusion of sources
3,892 records identified through database screening (ERIC: 747; PsycINFO: 1,237; Embase: 1,191; PubMed: 717)
3,603 duplicates removed and excluded records based on title and/or abstract
294 full-text sources assessed for eligibility
165 full-text sources excluded with reasons (Tab. S2)
129 sources identified through database screening
129 sources identified by hand search
258 sources included

The included sources describe 21 different scoring methods for single-choice items. In the following subsections, all scoring methods are described with their corresponding scoring formulas for calculating examination results as absolute scores (*S*). In addition, an overview with the respective scoring results for individual items as well as alternative names used in the literature is presented in Table 2. All abbreviations used throughout this review are shown at the end of this review.

Table 2. Identified scoring methods and algorithms for single-choice items.

Method Number and Sources	**Scoring Method**	**Algorithm**
1 [5, 6, 16-172]	• *0-1 score* [167] • *Zero-one scoring* [146] • *Binary scoring* [146] • *Dichotomous scoring* [105, 114] • *All-or-none scoring* [166]	$f = 1$ (if $i = 1$) $f = 0$ (otherwise)

	- *Number-right (NR) scoring* [6, 20, 21, 24, 25, 27, 29-31, 37, 39, 50, 54, 56, 66, 67, 71, 73, 76, 79, 80, 85, 87, 95, 97, 99, 100, 111, 128, 132, 140, 145, 147, 153, 157, 160, 164] - *Number of right (NR) rule* [139] - *No. right score (No Rt)* [42] - *NC scoring* [144] - *Rights score* [72, 82, 92] - *R method* [24, 29, 39] - *Number correct scoring* [101, 106, 114, 124, 138, 151, 154, 155] - *Percentage-correct scoring* [165] - *Raw score* [44-46, 48, 51, 54, 57, 68, 86, 102, 118, 125, 131, 135] - *Score = rights* [23, 24] - *Uncorrected score* [91, 122, 137] - *Conventional scoring* [98] - *Rights-only score* [62, 87] - *3 right minus 0 wrong* [17]	
2 [37, 41, 46, 53, 58, 60, 65, 67, 79-81, 87, 91, 98, 111, 122, 137, 173-180]	- *Formula scoring* [67] - *Omission-formula scoring* [79] - *Omit-correction* [180] - *Positive scoring rule* [139] - *Adjusted score* [91]	f = 1 (if i = 1) f = 1 / n (if o = 1) f = 0 (otherwise)
3 [154]	*Fair penalty* [154]	f = 1 (if i = 1)

		f = 0 (if o = 1) f = 1 – 1 / n (otherwise)
4 [181]		f = 1 / (n – 1) (if i = 1) f = 0 (if o = 1) f = 0 (otherwise)
5 [80, 100, 182]		f = 1 (if i = 1) f = 0 (if o = 1) f = -1 / [2 (n – 1)] (otherwise)
6 [5, 23-29, 34, 37, 44, 46, 48, 50, 51, 53-57, 59-62, 64, 65, 67, 68, 70, 71, 74, 75, 79-81, 85-88, 91, 92, 98-101, 105, 106, 111, 113, 120, 122, 124-126, 128, 130, 134, 135, 137-139, 144, 145, 160, 169, 173-179, 182-225]	• *Formula scoring* [67, 85, 92, 101, 128, 160, 225] • *Conventional-formula scoring* [79] • *Conventional correction-for-guessing formula* [80, 213] • *Conventional correction formula* [201] • *‚Neutral‘ counter-marking* [88] • *CG scoring* [144] • *Negative marking* [130, 145] • *Logical marking* [130] • *Correction for blind guessing (cfbg)* [135] • *Correction for guessing (CFG) formula* [50, 51, 56, 57, 62, 71, 86, 87, 99, 101, 105, 106, 113, 122, 124, 137, 176, 179, 195, 199, 204, 223] • *Correction for chance formula* [56, 87, 174, 188] • *Discouraging guessing* [138] • *Rights minus wrongs correction* [98]	f = 1 (if i = 1) f = 0 (if o = 1) f = -1 / (n – 1) (*otherwise*)

	- *Corrected score* [37, 48, 55, 59, 68, 91] - *Classical score* [207] - *Mixed rule* [139]	
7 [226]		$f = 1 / (n - 1)$ (if $i = 1$) $f = 0$ (if $o = 1$) $f = -1 / (n - 1)$ (*otherwise*)
8 [41]		$f = (n - 1) / n$ (if $i = 1$) $f = 0$ (if $o = 1$) $f = -1 / n$ (*otherwise*)
9 [6, 48, 62, 88, 224, 227, 228]	- *3 right – wrong* [6] - *Negative marking* [228]	$f = 1$ (if $i = 1$) $f = 0$ (if $o = 1$) $f = -1 / 3$ (*otherwise*)
10[a] [229]		$f = 1$ (if $i = 1$) $f = 0$ (if $o = 1$) $f = -0.48$ (*otherwise*
11 [18, 23, 41, 62, 69, 224, 229-234]		$f = 1$ (if $i = 1$) $f = 0$ (if $o = 1$) $f = -0.5$ (*otherwise*)
12[a] [229, 231]		$f = 1$ (if $i = 1$) $f = 0$ (if $o = 1$) $f = -0.6$ (*otherwise*)
13 [4, 6, 16-19, 21-25, 29-33, 38, 39, 42, 43, 45, 49, 52, 55, 69, 72, 76, 82, 110, 130, 132, 140, 143, 154, 157, 164, 172,	- *Formula scoring* [157, 164] - *Correct-minus-incorrect score* [267] - *C-I score* [132] - *R – W method* [23, 24, 29, 30, 32, 38, 39, 42, 76, 243, 245, 246, 249, 259] - *Number right minus number wrong method* [39, 45] - *Right-minus-wrong method* [6, 21, 23, 25, 30,	$f = 1$ (if $i = 1$) $f = 0$ (if $o = 1$) $f = -1$ (*otherwise*)

190, 193, 215, 216, 219, 229, 232, 233, 235-267]	31, 42, 72, 82, 236, 244, 247] • *Rights minus wrongs method* [29, 253, 254, 256, 258] • *Right – wrong* [266] • *T – F formula* [260] • *Guessing penalty* [154] • *Correction-for-guessing* [76, 128] • *Negative marking* [140] • *Logical marking* [130] • *1 right minus 1 wrong* [17] • *Penal guessing formula* [55] • *Corrected score* [265]	
14^a [249, 268]		$f = 1$ (if $i = 1$) $f = 0.7$ (if $o = 1$) $f = -1$ (*otherwise*)
15^a [186]		$f = 1$ (if $i = 1$) $f = 0.7$ (if $o = 1$) $f = -1.1$ (*otherwise*)
16 [20]		$f = 1$ (if $i = 1$) $f = 0$ (if $o = 1$) $f = -n / (n – 1)$ (*otherwise*)
17^a [203, 259]		$f = 1$ (if $i = 1$) $f = 0$ (if $o = 1$) $f = -1.5$ (*otherwise*)
18^a [203]		$f = 1$ (if $i = 1$) $f = 0$ (if $o = 1$) $f = -1.8$ (*otherwise*)
19 [6, 17, 20, 21, 49, 75,	• *Right – 2 wrong* [6] • *1 right minus 2 wrong* [17]	$f = 1$ (if $i = 1$) $f = 0$ (if $o = 1$)

203, 253, 268-270]	• *Rights minus two times wrongs* [253] • *r-2w* [253]	f = -2 / (n – 1) (*otherwise*)
20[a] [17, 41]	• *1 right minus 3 wrong* [17]	f = 1 (if i = 1) f = 0 (if o = 1) f = -3 (*otherwise*)
21[b] [259]		f = 1 (if i = 1) f = 0 (if o = 1) f = -62 / 38 (if i = 0 and t_m = 0) f = -38 / 62 (if i = 0 and t_m = 1)

f = resulting score per item, i = 1 if the item was marked correctly, otherwise i = 0; n = number of answer options per item ($n \geq 2$); o = 1 if the item was omitted, otherwise o = 0; t_m = 1 if the statement is true, otherwise t_m = 0. [a]Only described for n = 2. [b]Only described for single true-false items.

Scoring Methods without Malus Points (0 to a Maximum of +1 Point per Item)

Method 1: One credit point is awarded for a correct response. Therefore, the examination result as absolute score (S) corresponds to the number of correct responses (R). No points are deducted for incorrect responses (W). The formula is $S = R$.

Method 2: One credit point is awarded for a correct response. In addition, 1 / n credit points per item are awarded for each omitted item (O). No points are deducted for incorrect responses. The formula is $S = R + O / n$. This scoring method was first described by *Lindquist* [37] in 1951.

Method 3: One credit point is awarded for a correct response. For incorrect responses,

$1 – 1 / n$ credit points are awarded. The formula is $S = R + (1 – 1 / n) W$. This scoring method was first described by Costagliola et al. [154] in 2007 and named *fair penalty* by the authors. However, the term *penalty* is misleading because no points are deducted in case of incorrect responses.

Method 4: For each correct response, $1 / (n – 1)$ credit points are awarded. Omitted items and incorrect responses do not affect the score. The formula is $S = R / (n – 1)$. For example, 1 credit point is awarded for a correct response on single-choice items with $n = 2$ (ie, alternate-choice items, single true-false items) but only 0.25 credit points are awarded for a correct response on best-answer items with $n = 5$. This scoring method was first described by *Foster and Ruch* [181] in 1927.

Scoring Methods with Malus Points (Maximum -1 to +1 Point per Item)

Method 5: One credit point is awarded for a correct response. For incorrect responses, $1 / [2 (n – 1)]$ points are deducted. The formula is $S = R – W / [2 (n – 1)]$. This scoring method was first described by *Little* [182] in 1962.

Method 6: One credit point is awarded for a correct response. For incorrect responses, $1 / (n – 1)$ points are deducted. The formula is $S = R – W / (n – 1)$. This scoring method was first described by *Holzinger* [183] in 1924. For items with $n = 2$, methods 6 and 13 result in identical scores; for items with $n = 4$, methods 6 and 9 result in identical scores.

Method 7: For each correct response, $1 / (n – 1)$ credit points are awarded. For an incorrect response, $1 / (n – 1)$ points are deducted. The formula is $S = (R – W) / (n – 1)$. This scoring method was first described by *Petz* [226] in 1978.

Method 8: For each correct response, $(n – 1) / \text{n}$ credit points are awarded. For an incorrect response, $1 / n$ points are deducted. Omissions do not affect the score. The formula is

$S = [(n - 1) R] / n - W / n$. As a result, examinees achieve only 0.5 credit points for each correct response on single-choice items with $n = 2$ and 0.8 credit points for each correct response on best-answer items with $n = 5$. This scoring method was first described by *Guilford* [41] in 1954.

Method 9: One credit point is awarded for a correct response. For incorrect responses, 1 / 3 points are deducted. The formula is $S = R - 1 / 3 W$. Originally, this scoring method was described by *Paterson and Langlie* [6] in 1925 with the formula $S = 3 R - W$ for items with $n = 2$ only. Later, the scoring method was also described for single-choice items with more answer options [88, 203]. For items with $n = 4$, methods 6 and 9 give identical results.

Method 10: One credit point is awarded for a correct response. For incorrect responses, 0.48 points are deducted. The formula is $S = R - 0.48 W$. This scoring method was first described by *Gupta and Penfold* [229] in 1961 for single-choice items with $n = 2$.

Method 11: One credit point is awarded for a correct response. Half a point is deducted for incorrect responses. The formula is $S = R - 0.5 W$. This scoring method was first described in 1924 by *Brinkley* [18] and *Asker* [230] for single-choice items with $n = 2$, but was later also used for single-choice items with more answer options.

Method 12: One credit point is awarded for a correct response. For incorrect responses, 0.6 points are deducted. The formula is $S = R - 0.6 W$. This scoring method was first described by *Gupta* [231] in 1957 for single-choice items with $n = 2$.

Method 13: One credit point is awarded for a correct response. One point is deducted for incorrect responses. The formula is $S = R - W$. For items with $n = 2$, methods 6 and 13 result in identical scores. This scoring method was first described by *McCall* [4] in 1920 for single-choice items with $n = 2$, but was later also used for single-choice items with more answer options.

Method 14: This scoring method results in 1 credit point for a correct response, 0.7 credit points for an omitted item, and -1 point for an incorrect response. The formula is $S = R + 0.7O - W$. This scoring method was first described by *Staffelbach* [268] in 1930 for single-choice items with $n = 2$.

Scoring Methods with Malus Points (Maximum -3 to +1 Points per Item)

Method 15: This scoring method results in 1 credit point for a correct response, 0.7 credit points for an omitted item, and -1.1 points for an incorrect response. The formula is $S = R + 0.7O - 1.1W$. This scoring method was first described by *Kinney and Eurich* [186] in 1933 for items with $n = 2$.

Method 16: One credit point is awarded for a correct response. For an incorrect response, $n / (n - 1)$ points are deducted. The formula is $S = R - n\,W / (n - 1)$. This scoring method was first described by Miller [20] in 1925. For items with $n = 2$, methods 16 and 19 result in identical scores.

Method 17: For an incorrect response, 1.5 times as many points are deducted as credit points are awarded for a correct response. The original scoring formula is $S = 2\,R - 3\,W$. If a maximum of 1 credit point is awarded per item, 1 credit point is awarded for a correct response and 1.5 points are deducted for an incorrect response. This results in the following scoring formula: $S = R - 1.5\,W$. This scoring method was first described by *Cronbach* [259] in 1942 for items with $n = 2$.

Method 18: One credit point is awarded for a correct response. For an incorrect response, 1.8 points are deducted. The scoring formula is $S = R - 1.8\,W$. This scoring method was first described by *Lennox* [203] in 1967 for items with $n = 2$.

Method 19: One credit point is awarded for a correct response. For an incorrect response, 2 / (*n* – 1) points are deducted. The formula is $S = R - 2\ W / (n - 1)$. This scoring method was first described by *Gates* [269] in 1921 with the scoring formula $S = R - 2\ W$ for items with $n = 2$. Later, the scoring formula was also described for single-choice items [203, 270]. In case of items with $n = 2$, methods 16 and 19 result in identical scores.

Method 20: One credit point is awarded for a correct response. Three points are deducted for an incorrect response. The formula is $S = R - 3\ W$. This method was first described by *Wood* [17] in 1923 for items with $n = 2$.

Specific Scoring Methods for Single True-false Items

Method 21: One credit point is awarded for correctly identifying the statement of true-false single items as true or false. If the statement presented is marked incorrectly, 62/38 points are deducted on true statements (W_t, incorrectly marked as false), but only 38/62 points are deducted on false statements (W_f, incorrectly marked as true). The scoring formula is $S = R - 62/38\ W_t - 38/62\ W_f$. This scoring method was first described by Cronbach [259] in 1942 for single true-false items and differentiates in the scoring of incorrectly marked true/false statements.

Expected Chance Scores of the Identified Scoring Methods

The *expected chance scores* of examinees without any knowledge ($k = 0$) varies between -1 and +0.75 credit points per item for single-choice items with $n = 2$. For single-choice items with $n = 5$, *expected chance scores* show a larger variability. Here, the *expected chance scores* vary between -2.2 and +0.84 credit points per item, depending on the selected scoring method. A detailed list is shown in Table 3.

Table 3. Overview of scoring results for single-choice items with $n = 2$ or $n = 5$ answer option.

Method Number	**Scoring Formula**	***n* = 2**			***n* = 5**		
		Credit for incorrect responses[a]	**Credit for correct responses** [b]	***Expected Chance Score***	**Credit for incorrect responses**[a]	**Credit for correct responses** [b]	***Expected Chance Score***
1	$S = R$	0	1	0.50	0	1	0.20
2	$S = R + O / n$	0	1	0.50	0	1	0.20
3	$S = R + (1 – 1 / n) W$	0.50	1	0.75	0.80	1	0.84
4	$S = R / (n – 1)$	0	1	0.50	0	0.25	0.05
5	$S = R – W / [2 (n – 1)]$	-0.50	1	0.25	-1 / 8	1	0.10
6	$S = R – W / (n – 1)$	-1	1	0.00	-0.25	1	0.00
7	$S = (R – W) / (n – 1)$	-1	1	0.00	-0.25	0.25	0.15
8	$S = [(n – 1) / n] R – W / n$	-0.50	0.50	0.00	-0.20	0.80	0.00
9	$S = R – (1 / 3) W$	-1 / 3	1	1 / 3	-1 / 3	1	-2 / 30
10	$S = R –$	-0.48	1	0.26	-0.48	1	-23 / 125

	$0.48\,W$						
11	$S = R - 0.5\,W$	-0.5	1	0.25	-0.5	1	-0.20
12	$S = R - 0.6\,W$	-0.6	1	0.20	-0.6	1	-0.28
13	$S = R - W$	-1	1	0.00	-1	1	-0.60
14	$S = R + 0.7\,O - W$	-1	1	0.00	-1	1	-0.60
15	$S = R + 0.7\,O - 1.1\,W$	-1.10	1	-0.05	-1.10	1	-0.68
16	$S = R - n\,W / (n - 1)$	-2	1	-0.50	-1.25	1	-0.80
17	$S = R - 1.5\,W$	-1.5	1	-0.25	-1.5	1	-1.00
18	$S = R - 1.8\,W$	-1.8	1	-0.40	-1.8	1	-1.24
19	$S = R - 2\,W / (n - 1)$	-2	1	-0.50	-0.5	1	-0.20
20	$S = R - 3\,W$	-3	1	-1.00	-3	1	-2.20
21	$S = R - (62/38)\,W_t - (38/62)\,W_f$	-62 / 38 or -38 / 62	1	N/A[c]	-62 / 38 or -38 / 62	1	N/A[c]

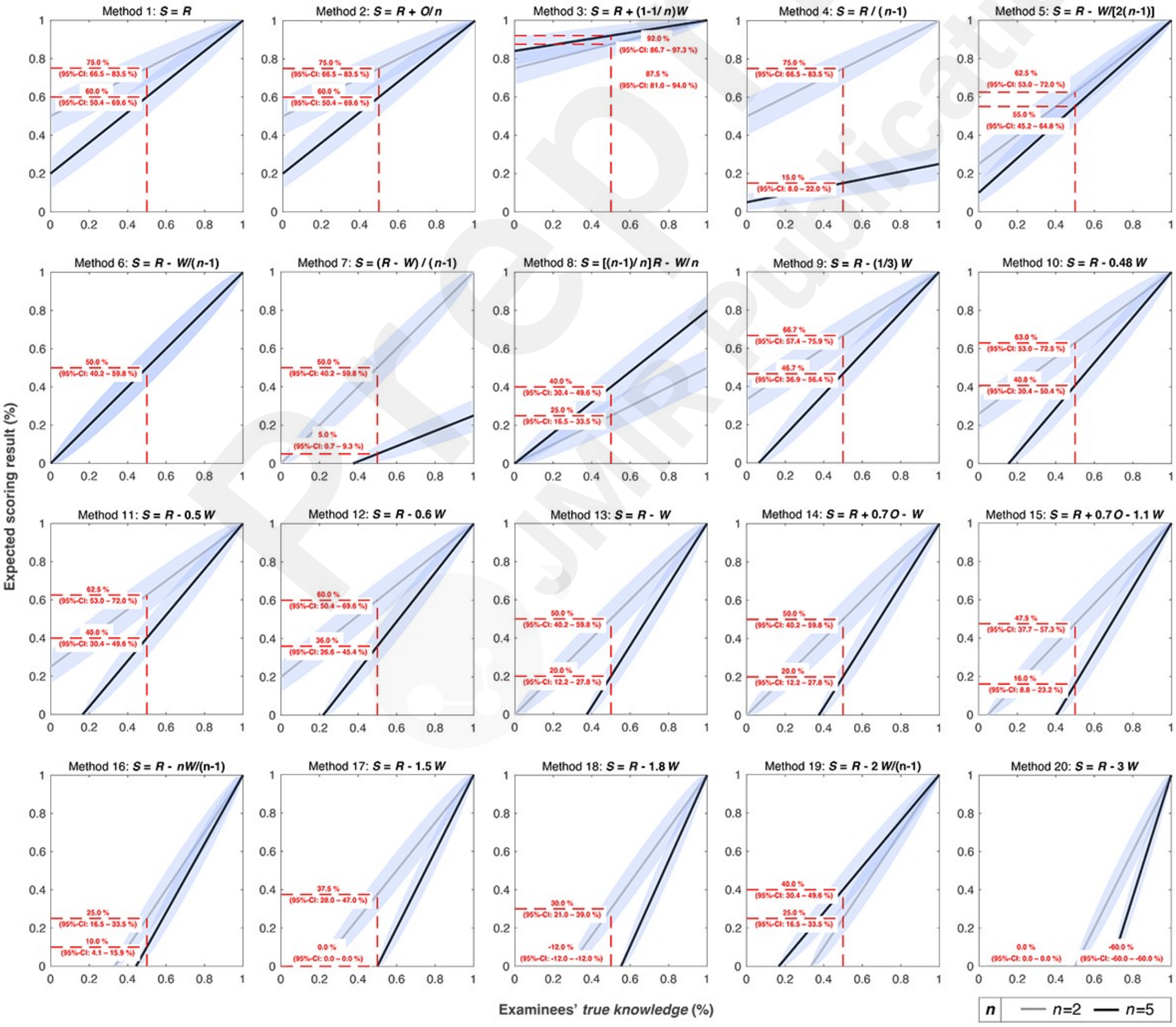
n = number of answer options per item ($n \geq 2$); O = number of omitted items;

S = examination result as absolute score; *W* = number of incorrect responses; W_f = number of true statements incorrectly marked as false; W_t = false statements incorrectly marked as true. [a]*R* = 0, *O* = 0, *W* = 1. [b]*R* = 1, *O* = 0, *W* = 0. [c]*Expected chance scores* were not calculated for method 21, since these depend on the proportion of true-false items with correct or incorrect statements.

Relation Between Examinees' *True Knowledge* and the Expected Scoring Results

The relation between examinees' *true knowledge* and expected scoring results for single-choice items with *n* = 2 and *n* = 5 are shown in Figure 4 (a high resolution image is shown in Multimedia Appendix 3). For all identified scoring methods, there is a linear relation between examinees' *true knowledge* and the expected scoring results. However, some scoring methods (ie, methods 4 and 7) award less than one point for correctly marked items if there are more than two answer options (*n* > 2). One further method (method 8) awards less than one point for correctly marked items regardless of the number of answer options, so the maximum score for these scoring methods might be less than 100%. Depending on the scoring method and the number of answer options, the y-axis intercepts (*expected chance scores*, *k* = 0) and the slopes differ. A low *expected chance score* results in a wide range of examination results which differentiate different examinees' knowledge levels (ranging from the *expected chance score* as the lower limit, to the maximum score as the upper limit). Only for methods 6 and 8 as well as method 7 in the case of *n* = 2, the line starts from the pole (ie, examinees without any knowledge [*k* = 0] achieve an examination result of 0%). Only for method 6, the relation between examinees' *true knowledge* and the expected scoring results is independent of the number of answer options per item.

Figure 4. Relation between examinees' *true knowledge* (%) and the expected scoring results for examinations with 100 single-choice items (either n = 2 or n = 5 answer options per item). In each case, the expected scoring result at 50% *true knowledge* is shown with the associated 95% confidence interval. Method 21 is not shown because the relation depends on the proportion of single true-false items with true or false statements. O = number of omitted items (O = 0); R = number of correct responses; S = examination result as absolute scores (max. up to 100 points); W = number of incorrect responses. Please see Multimedia Appendix 3 for high resolution image.



Discussion

In this review, a total of 21 scoring methods for single-choice items could be identified. The majority of identified scoring methods is based on theoretical considerations or empirical findings, while others have been arbitrarily determined. Although some methods were only described for certain item types (ie, single-choice items with *n* = 2), most of them might also be used for scoring items with more answer options. However, one method is suitable for scoring single true-false items only.

Principal Findings

All scoring methods have in common that omitted items do not result in any credit deduction. Some scoring methods even award a fixed amount of 0.7 points on omitted items (methods 14 and 15), which is, however, lower than the full credit for a correct response, or the score to be achieved on average by guessing (1 / *n*, method 2).

For the identified scoring methods, the possible scores range from a maximum of -3 to +1 points. A correctly marked item is usually scored with one full point (1 credit point). Exceptions to this are three scoring methods which only award 1 credit point in case of single-choice items with *n* = 2 (methods 4 and 7) or which never award 1 credit point (method 8). These scoring methods are questionable because as the number of answer options increases, the guessing probability decreases. Also, a differentiation between examinees' marking on true and false statements (method 21) is not justified, since the importance of correctly identifying true statements (ie, correctly marking the statement as true) and false statements (ie, correctly marking the statement as false) is likely to be considered equivalent in the context of many examinations.

With the exception of method 6, the relation between examinees' *true knowledge* and the resulting examination scores depends on the number of answer options per item (*n*).

Therefore, *n* must usually be taken into account when examination scores are interpreted.

Examinations are designed to determine examinees' knowledge as well as to decide whether the examinees pass or fail in summative examinations. It can be generally assumed that examinees must perform at least 50% of the expected performance to receive at least a passing grade [271]. If examinees are to be tested on a *true knowledge* of 50%, adjusted pass marks must be applied depending on the scoring method used and the number of answer options per item. The theoretical considerations show that for an examination testing for 50% *true knowledge*, a pass mark of 0% or even negative scoring results might be appropriate, while other scoring methods would require pass marks up to 92%. Consequently, the examination's pass mark must be considered or adjusted when selecting a suitable scoring method. However, the pass mark might be fixed due to local university or national guidelines resulting in a limited number of suitable scoring methods.

Correction for Guessing

To account for guessing in case of single true-false items, the scoring formula *R* – *W* (method 13) was originally propagated in the literature, where the number of incorrect responses is subtracted from the number of correct responses [4]. Since its first publication in 1920, this scoring method has been frequently criticized: the main criticism is that this scoring method assumes examinees to either have complete knowledge ($k = 1$) or to guess blindly ($k = 0$). However, especially in the context of university examinations, examinees are assumed to have at least some partial knowledge. Furthermore, the scoring method assumes that incorrect responses are exclusively the result of guessing. No differentiation is made between incorrect responses due to blind guessing (ie, complete lack of knowledge), informed guessing (ie, guessing with partial knowledge and remaining uncertainty) or other reasons (eg, transcription errors introduced when transferring markings to the answer sheet) despite complete knowledge. Due to the 50% guessing probability in case of *alternate-choice items* or *single true-false items*, it is assumed that for each incorrectly guessed response (*W*) one item is also marked correctly by guessing on average, so that the corrected result is obtained by the scoring formula *R* – *W*. Especially in the case of partial knowledge, examinees' marking behavior not only depends on their actual knowledge but also on their individual personality (eg, risk-seeking behavior) [272]. Consequently, the construct validity of examinations must be questioned when using the scoring formula *R* – *W*. Another criticism is that a correction by awarding malus points does not change the relative ranking of the results of different examinees if all examinees have sufficient time to take the examination and all items are answered [44, 46].

Therefore, alternative scoring methods and scoring formulas emerged in addition to the already discussed scoring formula *R* – *W*. In this context, the literature often refers to *formula scoring*. However, the term *formula scoring* is not used uniformly: on the one hand, it is used as a general umbrella term for various scoring methods to correct for the guessing probability. On the other hand, the term is used to refer to specific scoring methods (methods 2, 6, and 13). Using method 2, examinees receive 1 / *n* points for each omitted item. This corresponds to the number of points they would have scored on average by blindly guessing. Method 6 is a generalization of the scoring formula *R* – *W* for variable numbers of answer options. In case of *n* answer options, there are *n* – 1 times as many incorrect answer options as correct answer options and it is assumed that for each incorrectly guessed response (*W*) also *W* / (*n* – 1) items are marked correctly by guessing on average. Therefore, the corrected score is given by the scoring formula *R* – *W* / (*n* – 1). Consequently, methods 6 and 13 yield identical scoring results in case of items with *n* = 2.

Strengths and Limitations

So far, the relation between examinees' *true knowledge* and the expected scoring result for single-choice items has been shown only for a small number of scoring methods [273]. Therefore, a systematic literature search was conducted in several databases as part of this review. As a result, a large number of different scoring methods have been identified and were compared in this review assisting (medical) educators in gaining a comprehensive overview and to allow for informed decisions regarding the scoring of single-choice items. However, limitations are also present: First, a number of assumptions (eg, equal difficulty of items and answer options, absence of cues) were required for simplification of the calculations and comparisons. However, these assumptions are likely to be violated in real examinations [15, 274-276]. Second, calculations are based on classical test theory assumptions and did not employ item response theory models which might yield different results. Third, databases were already searched in September 2020 and potentially eligible sources published thereafter might not be included in this review. However, single-choice items have been used in examinations for over 100 years and further scoring methods are unlikely to have emerged in the past 2 years.

Comparison to Prior Work

Even though some of the identified scoring methods might also be applied to other item formats (eg, *multiple-select items*), the presented equation for the calculation of the expected scoring result is limited to single-choice items. Analogous calculations for items in multiple-select multiple-choice formats with (eg, *Pick-N* items) or without (*Multiple-True-False* items) mutual stochastic dependence have already been described in the literature [11, 14].

Practical Implications

In practice, the evaluation of a multiple-choice examination should be based on an easy-to-calculate scoring method that allows for a transparent credit awarding and is accepted by jurisdiction. In this regard, scoring methods with malus points (ie, methods 5-21) may not be accepted by national jurisdiction in certain countries (eg, Germany) [277]. Furthermore, it does not seem reasonable to discourage examinees from marking an item by awarding malus points for the reasons already mentioned. Therefore, only four of the presented scoring methods can be used versatile. Furthermore, it seems inconclusive to reward partial credit on incorrect responses or to refrain from awarding 1 credit point for correct responses in case of items with more than two answer options ($n > 2$). As a result, only a dichotomous scoring method (1 credit point for a correct response, 0 points for an incorrect response or omitted items) is recommended. Within the context of this review, the outlined scoring method is referred to as method 1.

The scoring of examinations with different item types, item formats, or items containing a varying number of answer options within a single examination is more complicated. Here, the individual examination sections would have to be evaluated separately or the credit resulting from the respective item type or item format would have to be corrected in order to enable a uniform pass mark. For example, in the single-choice format, credit points resulting from items with $n = 2$ would have to be reduced to compensate for the higher guessing probability compared to items with $n = 5$ (ie, 50% vs. 20% guessing probability).

Conclusion

Single-response items only allow clearly correct or incorrect responses from examinees. Consequently, the scoring should also be dichotomous and result in either 0 points (incorrect response) or 1 credit point (correct response) per item. Due to the possibility of guessing, scoring results cannot be equated with examinees' *true knowledge*. If (medical) educators interpret scoring results and determine suitable pass marks, the *expected chance score* must be taken into account, which in the proposed dichotomous scoring methods depends on the number of answer options per item.

Acknowledgements

The authors acknowledge support by the Open Access Publication Funds of Göttingen University. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Data Availability

All data generated or analyzed during this study are included in this published article and its supplementary information files.

Authors' Contributions

AFK and PK contributed to the study's conception and design, performed the literature search and data extraction, and drafted the manuscript. PK performed statistical analyses. All authors interpreted the data, critically revised the manuscript, and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Abbreviations

k: examinees’ true knowledge

n: number of answer options per item

O: number of omitted items

R: number of correct responses

S: examination result as absolute scores

W: number of incorrect responses

W_f: number of true statements incorrectly marked as false

W_t: number of false statements incorrectly marked as true

Multimedia Appendix 1: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews Checklist.

Multimedia Appendix 2: Excluded sources after screening of full-texts.

Multimedia Appendix 3. High resolution version of Figure 4.

References

1. Krebs R. Anleitung zur Herstellung von MC-Fragen und MC-Prüfungen für die ärztliche Ausbildung [Instructions for preparing multiple choice items and multiple choice examinations in medical education]. Bern, Switzerland: Department for Assessment and Evaluation (AAE), Institute for Medical Education, University of Bern; 2004.

2. Ebel RL. Proposed solutions to two problems of test construction. J Educ Meas 1982;19(4):267-78. doi:10.1111/j.1745-3984.1982.tb00133.x

3. Kelly FJ. The Kansas silent reading test. J Educ Psychol 1916;7(2):63-80. doi:10.1037/h0073542

4. McCall WA. A new kind of school examination. J Educ Res 1920;1(1):33-46. doi:10.1080/00220671.1920.10879021

5. Ruch GM, Stoddard GD. Comparative reliabilities of five types of objective examinations. J Educ Psychol 1925;16(2):89-103. doi:10.1037/h0072894

6. Paterson DG, Langlie TA. Empirical data on the scoring of true-false tests. J Appl Psychol 1925;9(4):339-48. doi:10.1037/h0069813

7. Lindner MA, Strobel B, Köller O. Multiple-Choice-Prüfungen an Hochschulen? Ein Literaturüberblick und Plädoyer für mehr praxisorientierte Forschung [Are multiple-choice exams useful for universities? A literature review and argument for a more practice oriented research]. Z Pädagog Psychol 2015;29(3-4):133-49. doi:10.1024/1010-0652/a000156

8. Mathysen DGP, Aclimandos W, Roelant E, Wouters K, Creuzot-Garcher C, Ringens PJ, et al. Evaluation of adding item-response theory analysis for evaluation of the European Board of Ophthalmology Diploma examination. Acta Ophthalmologica 2013;91(7):e573-e7. PMID:23927770 doi:10.1111/aos.12135

9. Rutgers DR, van Raamt F, van der Gijp A, Mol C, Ten Cate O. Determinants of difficulty and discriminating power of image-based test items in postgraduate radiological

examinations. Academic Radiology 2018;25(5):665-72. PMID:29198947 doi:10.1016/j.acra.2017.10.014

10. Hubbard JP. Measuring medical education: The tests and test procedures of the National Board of Medical Examiners. Philadelphia, PA: Lea and Febiger; 1971. ISBN:9780812103656

11. Schmidt D, Raupach T, Wiegand A, Herrmann M, Kanzow P. Relation between examinees’ true knowledge and examination scores: systematic review and exemplary calculations on Multiple-True-False items. Educ Res Rev 2021;34:Article 100409. doi:10.1016/j.edurev.2021.100409

12. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. Ann Intern Med 2018;169(7):467-73. PMID:30178033 doi:10.7326/M18-0850

13. Albanese MA, Sabers DL. Multiple true-false items: A study of interitem correlations, scoring alternatives, and reliability estimation. J Educ Meas 1988;25(2):111-23. doi:10.1111/j.1745-3984.1988.tb00296.x

14. Schmidt D, Raupach T, Wiegand A, Herrmann M, Kanzow P. Relation between examinees’ true knowledge and examination scores: systematic review and exemplary calculations on Pick-N items. Educ Res Rev 2022;37:Article 100483. doi:10.1016/j.edurev.2022.100483

15. Kanzow P, Schuelper N, Witt D, Wassmann T, Sennhenn-Kirchner S, Wiegand A, et al. Effect of different scoring approaches upon credit assignment when using Multiple True-False items in dental undergraduate examinations. Eur J Dent Educ 2018;22(4):e669-e78. PMID:29934980 doi:10.1111/eje.12372

16. Toops HA. Trade Tests in Education. New York, NY: Teachers College, Columbia University; 1921.

17. Wood BD. Measurement in Higher Education. New York, NY: Teachers College, Columbia University; 1923.

18. Brinkley SG. Values of new type examinations in the high school. With special reference to history. New York, NY: Teachers College, Columbia University; 1924.

19. Farwell HW. The new type examinations in Physics. School Soc 1924;19(481):315-22.

20. Miller GF. Formulas for scoring tests in which the maximum amount of chance is determined. Proc Okla Acad Sci 1925;5:30-42.

21. Boyd W. An exploration of the true-false method of examination. Forum Educ 1926;4:34-8.

22. Christensen AM. A suggestion as to correcting guessing in examinations. J Educ Res 1926;14(5):370-4. doi:10.1080/00220671.1926.10879703

23. Ruch GM, Degraff MH, Gordon WE, McGregor JB, Maupin N, Murdock JR. Objective examination methods in the social studies. Chicago, IL: Scott, Foresman and Company; 1926.

24. Wood BD. Studies of achievement tests. J Educ Psychol 1926;17(1):1-22. doi:10.1037/h0076061

25. Wood EP. Improving the validity of collegiate achievement tests. J Educ Psychol 1927;18(1):18-25. doi:10.1037/h0070659

26. Greene HA. A new correction for chance in examinations of alternate-response type. J Educ Res 1928;17(2):102-7. doi:10.1080/00220671.1928.10879818

27. Odell CW. Traditional examinations and new-type tests. New York: The Century; 1928.

28. Ruch GM, Charles JW. A comparison of five types of objective tests in elementary psychology. J Appl Psychol 1928;12(4):398-403. doi:10.1037/h0075108

29. Cocks AW. The Pedagogical Value of the True-False Examination. Baltimore, MD: Warwick and York; 1929.

30. Dunlap JW, De Mello A, Cureton EE. The effects of different directions and scoring methods on the reliability of a true-false test. School Soc 1929;30(768):378-82.

31. Hevner K. A method of correcting for guessing in true-false tests and empirical evidence in support of it. J Soc Psychol 1932;3(3):359-62. doi:10.1080/00224545.1932.9919159

32. Melbo IR. How much do students guess in taking true-false examinations? Educ Method 1932/33;12:485-7.

33. Hawkes HE, Lindquist EF, Mann CR. The Construction and Use of Achievement Examinations: A Manual for Secondary School Teachers. Bostan, MA: Houghton Mifflin; 1936.

34. Rinsland HD. Constructing Tests and Grading in Elementary and High School Subjects. New York, NY: Prentice-Hall; 1937.

35. Lord FM. Reliability of multiple-choice tests as a function of number of choices per item. J Educ Psychol 1944;35(3):175-80. doi:10.1037/h0061025

36. Engelhart MD. Suggestions for writing achievement exercises to be used in tests scored on the electric scoring machine. Educ Psychol Meas 1949;7:357-74. doi:10.1177/001316444700700301

37. Lindquist EF. Educational Measurement. Washington, DC: American Council on Education; 1951.

38. Heston JC. How to take a test. Oxford, UK: Science Research Associates; 1953.

39. Keislar ER. Test instructions and scoring method in true-false tests. J Exp Educ 1953;21(3):243-9. doi:10.1080/00220973.1953.11010457

40. Swineford F, Miller PM. Effects of directions regarding guessing on item statistics of a

multiple-choice vocabulary test. J Educ Psychol 1953;44(3):129-39. doi:10.1037/h0057890

41. Guilford JP. Psychometric Methods. New York, NY: McGraw-Hill; 1954.

42. Sherriffs AC, Boomer DS. Who is penalized by the penalty for guessing? J Educ Psychol 1954;45(2):81-90. doi:10.1037/h0053756

43. Davis FB. Use of correction for chance success in test scoring. Educ Meas 1959;52(7):279-80. doi:10.1080/00220671.1959.10882581

44. Hubbard JP, Clemans WV. Multiple-Choice Examinations in Medicine: A Guide for Examiner and Examinee. Philadelphia, PA: Lea and Febiger; 1961.

45. Durost WN, Prescott GA. Essentials of Measurement for Teachers. New York, NY: Harcourt, Brace & World; 1962.

46. Ebel RL. Measuring educational achievement. Englewood Cliffs, NJ: Prentice-Hall; 1965.

47. Mattson D. The effects of guessing on the standard error of measurement and the reliability of test scores. Educ Psychol Meas 1965;15(3):727-30. doi:10.1177/001316446502500305

48. Cooper B, Fox JM. Guessing in multiple-choice tests. Brit J Med Educ 1967;1(3):212-5. PMID:6080737 doi:10.1111/j.1365-2923.1967.tb01699.x

49. Lennox B. Multiple choice. Brit J Med Educ 1967;1(5):340-4. PMID:5583311 doi:10.1111/j.1365-2923.1967.tb01728.x

50. Gronlund NE. Constructing Achievement Tests. Englewood Cliffs, NJ: Prentice-Hall; 1968.

51. Sax G, Collet L. The effects of differing instructions and guessing formulas on reliability and validity. Educ Psychol Meas 1968;28(4):1127-36. doi:10.1177/001316446802800411

52. Macintosh HG, Morrison RB. Objective Testing. London, UK: University of London

Press; 1969. ISBN:9780340096437

53. Traub RE, Hambleton RK, Singh B. Effects of promised reward and threatened penalty on performance of a multiple-choice vocabulary test. Educ Psychol Meas 1969;29(4):847-61. doi:10.1177/001316446902900410

54. Cronbach LJ. Essentials of Psychological Testing. 3rd ed. New York, NY: Harper & Row; 1970. ISBN:9780063561267

55. Houston JG. The Principles of Objective Testing in Physics. London, UK: Heinemann Educational Books; 1970. ISBN:9780435674229

56. Gronlund NE. Measurement and Evaluation in Teaching. 2nd ed. New York, NY: Macmillan; 1971. ISBN:9780023481802

57. Lyman HB. Test scores and what they mean. 2nd ed. Englewood Cliffs, NJ: Prentice-Hall; 1971. ISBN:9780139037818

58. Brandenburg DC, Whitney DR. Matched pair true-false scoring: Effect on reliability and validity. J Educ Meas 1972;9(4):297-302. doi:10.1111/j.1745-3984.1972.tb00961.x

59. Campbell CVT, Milne WJ. The Principles of Objective Testing in Chemistry. London, UK: Heinemann Educational Books; 1972. ISBN:9780435645755

60. Ebel RL. Essentials of Educational Measurement. Englewood Cliffs, NJ: Prentice-Hall; 1972. ISBN:9780132859998

61. Fraser WG, Gillam JN. The Principles of Objective Testing in Mathematics. London, UK: Heinemann Educational Books; 1972. ISBN:9780435503307

62. Diamond J, Evans W. The correction for guessing. Rev Educ Res 1973;32(2):181-91. doi:10.3102/00346543043002181

63. Rust WB. Objective Testing in Education and Training. London, UK: Pitman; 1973. ISBN:9780273316640

64. Hill GC, Woods GT. Multiple True-False questions. Educ Chem 1974;11(3):86-7.

65. Abu-Sayf FK. Relative effectiveness of the conventional formula score. J Educ Res 1975;69(4):160-2. doi:10.1080/00220671.1975.10884861

66. Hakstian AR, Kansup W. A comparison of several methods of assessing partial knowledge in multiple-choice tests: II. testing procedures. J Educ Meas 1975;12(4):231-9. doi:10.1111/j.1745-3984.1975.tb01024.x

67. Lord FM. Formula scoring and number-right scoring. J Educ Meas 1975;12(1):7-11. doi:10.1111/j.1745-3984.1975.tb01003.x

68. Brown FG. Principles of Educational and Psychological Testing. 2nd ed. New York, NY: Holt, Rinehart and Winston; 1976. ISBN:9780030890512

69. Harden RM, Brown RA, Biran LA, Dallas Ross WP, Wakeford RE. Multiple choice questions: To guess or not to guess. Med Educ 1976;10(1):27-32. PMID:1263885 doi:10.1111/j.1365-2923.1976.tb00527.x

70. Albanese MA, Kent TH, Whitney DR. A comparison of the difficulty, reliability and validity of complex multiple choice, multiple response and multiple true-false items. Annu Conf Res Med Educ 1977;16:105-10. PMID:606061

71. Cross LH, Frary RB. An empirical test of Lord's theoretical results regarding formula scoring of multiple-choice tests. J Educ Meas 1977;14(4):313-21. doi:10.1111/j.1745-3984.1977.tb00047.x

72. Eakin RR, Long CA. Dodging the dilemma of true-false testing. Educ Psychol Meas 1977;37(3):659-63. doi:10.1177/001316447703700308

73. Lord FM. Optimal number of choices per item - a comparison of four approaches. J Educ Meas 1977;14(1):33-8. doi:10.1111/j.1745-3984.1977.tb00026.x

74. Reid F. An alternative scoring formula for multiple-choice and true-false tests. J Educ Res 1977;70(6):335-9. doi:10.1080/00220671.1977.10885018

75. Whitby LG. Marking systems for multiple choice examinations. Med Educ

1977;11(3):216-20. PMID:865344 doi:10.1111/j.1365-2923.1977.tb00596.x

76. Aiken LR, Williams EN. Effects of instructions, option keying, and knowledge of test material on seven methods of scoring two-option items. Educ Psychol Meas 1978;38(1):53-9. doi:10.1177/001316447803800108

77. Hubbard JP. Measuring Medical Education: The Tests and Test Procedures of the National Board of Medical Examiners. 2nd ed. Philadelphia, PA: Lea and Febiger; 1978. ISBN:9780812106251

78. Morgan MKM, Irby DM. Evaluating Clinical Competence in the Health Professions. St. Louis, MO: Mosby; 1978. ISBN:9780801634932

79. Abu-Sayf FK. Recent developments in the scoring of multiple-choice items. Educ Rev 1979;31(3):269-79. doi:10.1080/0013191790310308

80. Abu-Sayf FK. The scoring of multiple choice tests: a closer look. Educ Technol 1979;19(6):5-15.

81. Ebel RL. Essentials of Educational Measurement. 3rd ed. Englewood Cliffs, NJ: Prentice-Hall; 1979. ISBN:9780132860130

82. Hsu LM. A comparison of three methods of scoring true-false tests. Educ Psychol Meas 1979;39(4):785-90. doi:10.1177/001316447903900411

83. Newble DI, Baxter A, Elmslie RG. A comparison of multiple-choice tests and free-response tests in examinations of clinical competence. Med Educ 1979;13:263-8. PMID:470647 doi:10.1111/j.1365-2923.1979.tb01511.x

84. Skakun EN, Nanson EM, Kling S, Taylor WC. A preliminary investigation of three types of multiple choice questions. Med Educ 1979;13:91-6. PMID:431421 doi:10.1111/j.1365-2923.1979.tb00928.x

85. Bliss LB. A test of Lord’s assumption regarding examinee guessing behavior on multiple-choice tests using elementary school students. J Educ Meas 1980;17(2):147-53.

doi:10.1111/j.1745-3984.1980.tb00823.x

86. Ahmann JS, Glock MD. Evaluating Student Progress: Principles of Tests and Measurements. 6th ed. Boston, MA: Allyn and Bacon; 1981. ISBN:9780205065615

87. Hopkins KD, Stanley JC. Educational and Psychological Measurement and Evaluation. 6th ed. Englewood Cliffs, NJ: Prentice-Hall; 1981. ISBN:9780132362733

88. Anderson J. Hand-scoring of multiple choice questions. Med Educ 1983;17(2):122-33. PMID:6843390 doi:10.1111/j.1365-2923.1983.tb01111.x

89. Kolstad RK, Briggs LD, Bryant BB, Kolstad RA. Complex multiple-choice items fail to measure achievement. J Res Develop Educ 1983;17(1):7-11.

90. Kolstad RK, Wagner MJ, Kolstad RA, Miller EG. The failure of distractors on complex multiple-choice items to prevent guessing. Educ Res Quart 1983;8(2):44-50.

91. Nitko AJ. Educational Tests and Measurement: An Introduction. New York, NY: Harcourt Brace Jovanovich; 1983. ISBN:9780155209107

92. Angoff WH, Schrader WB. A study of hypotheses basic to the use of rights and formula scores. J Educ Meas 1984;21(1):1-17. doi:10.1111/j.1745-3984.1984.tb00217.x

93. Diekhoff GM. True-false tests that measure and promote structural understanding. Teach Psychol 1984;11(2):99-101. doi:10.1207/s15328023top1102_11

94. Kolstad RK, Kolstad RA. The construction of machine-scored examinations: MTF clusters are preferable to CMC items. Sci Paedagog Exp 1984;21(1):45-54.

95. Norcini JJ, Swanson DB, Grosso LJ, Shea JA, Webster GD. A comparison of knowledge, synthesis, and clinical judgment. Multiple-choice questions in the assessment of physician competence. Eval Health Prof 1984;7(4):485-99. PMID:10269331 doi:10.1177/016327878400700409

96. Kolstad RK, Kolstad RA. Multiple-choice test items are unsuitable for measuring the learning of complex instructional objectives. Sci Paedagog Exp 1985;22(1):68-76.

97. Norcini JJ, Swanson DB, Grosso LJ, Webster GD. Reliability, validity and efficiency of multiple choice question and patient management problem item formats in assessment of clinical competence. Med Educ 1985;19(3):238-47. PMID:4010571 doi:10.1111/j.1365-2923.1985.tb01314.x

98. Crocker LM, Algina J. Introduction to Classical and Modern Test Theory. Orlando, FL: Holt, Rinehart and Winston; 1986. ISBN:9780030616341

99. Jaradat D, Sawaged S. The subset selection technique for multiple-choice tests: an empirical inquiry. J Educ Meas 1986;23(4):369-76. doi:10.1111/j.1745-3984.1986.tb00256.x

100. Aiken LR. Testing with multiple-choice items. J Res Develop Educ 1987;20(4):44-58.

101. Friedman MA, Hopwood LE, Moulder JE, Cox JD. The potential use of the Discouraging Random Guessing (DRG) approach in multiple-choice exams in medical education. Med Teach 1987;9(3):333-41. PMID:3683144 doi:10.3109/01421598709034796

102. Carey LM. Measuring and Evaluating School Learning. Newton, MA: Allyn and Bacon; 1988. ISBN:9780205111091

103. Osterlind SJ. Constructing test items. Boston, MA: Kluwer Academic Publishers; 1989. ISBN:9789401069717

104. Richards BF, Philp EB, Philp JR. Scoring the Objective Structured Clinical Examination using a microcomputer. Med Educ 1989;23(4):376-80. doi:10.1111/j.1365-2923.1989.tb01563.x

105. Cangelosi JS. Designing Tests for Evaluating Student Achievement. White Plains, NY: Longman; 1990. ISBN:9780801302633

106. Popham WJ. Modern Educational Measurement: A Practitioner’s Perspective. 2nd ed. Needham Heights, MA: Allyn and Bacon; 1990. ISBN:9780135938980

107. Moussa MAA, Ouda BA, Nemeth A. Analysis of multiple-choice items. Comput Methods Programs Biomed 1991;34(4):283-9. PMID:1873997 doi:10.1016/0169-

2607(91)90113-8

108. Viniegra L, Jiménez JL, Pérez-Padilla JR. El desafío de la evaluación de la competencia clínica [The challenge of evaluating clinical competence]. Rev Invest Clin 1991;43(1):87-98. PMID:1866504

109. Harasym PH, Price PG, Brant R, Violato C, Lorscheider FL. Evaluation of negation in stems of multiple-choice items. Eval Health Prof 1992;15(2):198-200. doi:10.1177/016327879201500205

110. Nnodim JO. Multiple-choice testing in anatomy. Med Educ 1992;26(4):301-9. PMID:1630332 doi:10.1111/j.1365-2923.1992.tb00173.x

111. Budescu D, Bar-Hille M. To guess or not to guess: a decision-theoretic view of formula scoring. J Educ Meas 1993;30(4):277-91. doi:10.1111/j.1745-3984.1993.tb00427.x

112. Fajardo LL, Chan KM. Evaluation of medical students in Radiology: written testing using uncued multiple-choice questions. Invest Radiol 1993;28(10):964-8. PMID:8262753 doi:10.1097/00004424-199310000-00020

113. Gronlund NE. How to Make Achievement Tests and Assessments. 5th ed. Needham Heights, MA: Allyn and Bacon; 1993. ISBN:9780205148240

114. Haladyna TM, Downing SM. How many options is enough for a multiple-choice test item? Educ Psychol Meas 1993;53(4):999-1010. doi:10.1177/0013164493053004013

115. Harasym PH, Doran ML, Brant R, Lorscheider FL. Negation in stems of single-response multiple-choice items. Eval Health Prof 1993;16(3):342-57. doi:10.1177/016327879301600307

116. Pinckney BA, Borcher GM, Clemens ET. Comparative studies of true/false, multiple choice and multiple-multiple choice. NACTA 1993;37(1):21-4.

117. Wolf DF. A comparison of assessment tasks used to measure FL reading comprehension. Mod Lang J 1993;77(4):473-89. doi:10.1111/j.1540-4781.1993.tb01995.x

118. Bott PA. Testing and Assessment in Occupational and Technical Education. Meedham Heights, MA: Allyn and Bacon; 1995. ISBN:9780205168781

119. Downing SM, Baranowski RA, Grosso LJ, Norcini JJ. Item type and cognitive ability measured: The validity evidence for multiple true-false items in medical specialty certification. Appl Meas Educ 1995;8(2):187-207. doi:10.1207/s15324818ame0802_5

120. Linn RL, Gronlund NE. Measurement and Assessment in Teaching. 7th ed. Englewood Cliffs, NJ: Merril; 1995. ISBN:9780023482618

121. Lumley JSP, Craven JL. Introduction. MCQ's in anatomy: A self-testing supplement to essential anatomy. 3rd ed. New York, NY: Churchill Livingstone; 1996.

122. Nitko AJ. Educational Assessment of Students. 2nd ed. Englewood Cliffs, NJ: Prentice Hall; 1996. ISBN:9780023876516

123. Schuwirth LWT, van der Vleuten CPM, Donkers HHL. A closer look at cueing effects in multiple-choice questions. Med Educ 1996;30(1):44-9. PMID:8736188 doi:10.1111/j.1365-2923.1996.tb00716.x

124. Ben-Simon A, Budescu DV, Nevo B. A comparative study of measures of partial knowledge in multiple-choice tests. Appl Psychol Meas 1997;21(1):65-88. doi:10.1177/0146621697211006

125. Thorndike RM. Measurement and Evaluation in Psychology and Education. Upper Saddle River, NJ: Merrill; 1997. ISBN:9780132541787

126. Gronlund NE. Assessment of student achievement. Needham Heights, MA: Allyn and Bacon; 1998. ISBN:9780205268580

127. Harasym PH, Leong EJ, Violato C, Brandt R, Lorscheider FL. Cuing effect of "all of the above" on the reliability and validity of multiple-choice test items. Eval Health Prof 1998;21(1):120-33. PMID:10183336 doi:10.1177/016327879802100106

128. Agble PK. A psychometric analysis of different scoring strategies in statistics

assessment. Kent, OH: Kent State University; 1999.

129. Bandaranayake R, Payne J, White S. Using multiple response true-false multiple choice questions. Aust N Z J Surg 1999;69(4):311-5. PMID:10327124 doi:10.1046/j.1440-1622.1999.01551.x

130. Burton RF, Miller DJ. Statistical modelling of multiple-choice and true/false tests: ways of considering, and of reducing, the uncertainties attributable to guessing. Assess Eval High Educ 1999;24(4):399-411. doi:10.1080/0260293990240404

131. Joint Committee on Standards for Educational and Psychological Testing of the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education. Standards for Educational and Psychological Testing. Washington, DC: American Educational Research Association; 1999. ISBN:9780935302257

132. Muijtjens AMM, Mameren HV, Hoogenboom RJI, Evers JLH, van der Vleuten CPM. The effect of a “don’t know” option on test scores: Number-right and formula scoring compared. Med Educ 1999;33(4):267-725. PMID:10336757 doi:10.1046/j.1365-2923.1999.00292.x

133. de Bruin WB, Fischhoff B. The effect of question format on measured HIV/AIDS knowledge: detention center teens, high school students, and adults. AIDS Educ Prev 2000;12(3):187-98. PMID:10926123

134. Linn RL, Gronlund NE. Measurement and Assessment in Teaching. 8th ed. Englewood Cliffs, NJ: Merril; 2000. ISBN:9780130983565

135. Beeckmans R, Eyckmans J, Janssens V, Dufranne M, Van de Velde H. Examining the yes/no vocabulary test: Some methodological issues in theory and practice. Lang Test 2001;18(3):235-74. doi:10.1177/026553220101800301

136. Blasberg R, Güngerich U, Müller Esterl W, Neumann D, Schappel S. Erfahrungen mit

dem Fragentyp „k aus n“ in Multiple-Choice-Klausuren [Experiences with item type "k from n" in multiple-choice-tests]. Med Ausbild 2001;18(S1):73-6.

137. Nitko AJ. Educational Assessment of Students. 3rd ed. Upper Saddle River, NJ: Merrill Prentice Hall; 2001. ISBN:9780130137081

138. Alnabhan M. An empirical investigation of the effects of three methods of handling guessing and risk taking on the psychometric indices of a test. Soc Behav Pers 2002;30(7):645-52. doi:10.2224/sbp.2002.30.7.645

139. Bereby-Meyer Y, Meyer J, Flascher OM. Prospect theory analysis of guessing in multiple choice tests. J Behav Decis Mak 2002;15(4):313-27. doi:10.1002/bdm.417

140. Burton RF. Misinformation, partial knowledge and guessing in true/false tests. Med Educ 2002;36(9):805-11. PMID:12354242 doi:10.1046/j.1365-2923.2002.01299.x

141. Griggs RA, Ransdell SE. Misconceptions tests or misconceived tests? In: Griggs RA, editor. Handbook for Teaching Introductory Psychology. Mahwah, NH: Lawrence Erlbaum Associates; 2002. p. 30-3.

142. Rahim SI, Abumadini MS. Comparative evaluation of multiple choice question formats. Introducing a knowledge score. Neurosciences (Riyadh) 2003;8(3):156-60. PMID:23649110

143. Anderson J. Multiple choice questions revisited. Med Educ 2004;26(2):110-3. PMID:15203517 doi:10.1080/0142159042000196141

144. Bradbard DA, Parker DF, Stone GL. An alternate multiple-choice scoring procedure in a macroeconomics course. Decis Sci J Innov Educ 2004;2(1):11-26. doi:10.1111/j.0011-7315.2004.00016.x

145. Burton RF. Multiple choice and true/false tests: Reliability measures and some implications of negative marking. Ass Eval High Educ 2004;29(5):585-95. doi:10.1080/02602930410001689153

146. Haladyna TM. Developing and Validating Multiple-Choice Test Items. 3rd ed. New York, NY: Routledge; 2004. ISBN:9780429238529

147. Burton RF. Multiple-choice and true/false tests: Myths and misapprehensions. Ass Eval High Educ 2005;30(1):65-72. doi:10.1080/0260293042003243904

148. Pamphlett R. It takes only 100 true-false items to test medical students: True or false? Med Educ 2005;27(5):468-72. PMID:16147803 doi:10.1080/01421590500097018

149. Swanson DB, Holtzman KZ, Clauser BE, Sawhill AJ. Psychometric characteristics and response times for one best-answer questions in relation to number and source of options. Acad Med 2005;80(S10):s93-s6. PMID:16199468 doi:10.1097/00001888-200510001-00025

150. MacCann R. The equivalence of online and traditional testing for different subpopulations and item types. Br J Educ Technol 2006;37(1):79-91. doi:10.1111/j.1467-8535.2005.00524.x

151. Shizuka T, Takeuchi O, Yashima T, Yoshizawa K. A comparison of three- and four-option English tests for university entrance selection purposes in Japan. Lang Test 2006;23(1):35-57. doi:10.1191/0265532206lt319oa

152. Swanson DB, Holtzman KZ, Allbee K, Clauser BE. Psychometric characteristics and response times for content-parallel extended-matching and one-best-answer items in relation to number of options. Acad Med 2006;81(S10):s52-s5. PMID:17001136 doi:10.1097/01.ACM.0000236518.87708.9d

153. Afolabi ERI. Effects of test format, self concept and anxiety on item response changing behaviour. Educ Res Rev 2007;2(9):255-8.

154. Costagliola G, Ferrucci F, Fuccella V, Oliveto R. eWorkbook: A computer aided assessment system. Int J Distance Educ Technologies 2007;5(3):24-41.

155. Downing SM, Yudkowsky R. Assessment in health professions education. New York,

NY: Routledge; 2009. ISBN:9780203880135

156. Tasdemir M. A comparison of multiple-choice tests and true-false tests used in evaluating student progress. J Instruct Psychol 2010;37(3):258-66.

157. Wakabayashi T, Guskin K. The effect of an “unsure” option on early childhood professionals’ pre- and post-training knowledge assessments. Am J Eval 2010;31(4):486-98. doi:10.1177/1098214010371818

158. Bayazit A, Aşkar P. Performance and duration differences between online and paper-pencil tests. Asia Pacific Education Review 2012;13(2):219-26. doi:10.1007/s12564-011-9190-9

159. Begum T. A guideline on developing effective multiple choice questions and construction of single best answer format. Journal of Bangladesh College of Physicians and Surgeons 2012;30(3):159-66. doi:10.3329/jbcps.v30i3.12466

160. Arnold MM, Higham PA, Martin-Luengo B. A little bias goes a long way: The effects of feedback on the strategic regulation of accuracy on formula-scored tests. J Exp Psychol Appl 2013;19(4):383-402. PMID:24341319 doi:10.1037/a0034833

161. Schaper ES, Tipod A, Ehlers JP. Use of key feature questions in summative assessment of veterinary medicine students. Ir Vet J 2013;66(1):Article 3. PMID:23497425 doi:10.1186/2046-0481-66-3

162. Simbak NB, Aung MMT, Ismail SB, Jusoh NBM, Ali TI, Yassin WAK, et al. Comparative study of different formats of MCQs: Multiple true-false and single best answer test formats, in a new medical school of Malaysia. Int Med J 2014;21(6):562-6.

163. Patil VC, Patil HV. Item analysis of medicine multiple choice questions (MCQs) for under graduate (3rd year MBBS) students. Res J Pharma Biol Chem Sci 2015;6(3):1242-51.

164. Ravesloot CJ, Van der Schaaf MF, Muijtjens AMM, Haaring C, Kruitwagen CLJJ,

Beek FJA, et al. The don’t know option in progress testing. Adv in Health Sci Educ 2015;20(5):1325-38. PMID:25912621 doi:10.1007/s10459-015-9604-2

165. Haladyna TM. Item analysis for selected response test items. In: Lane S, Raymond MR, Haladyna TM, editors. Handbook of test development. 2nd ed. New York, NY: Routledge; 2016.

166. Rush BR, Rankin DC, White BJ. The impact of item-writing flaws and item complexity on examination item difficulty and discrimination value. BMC Med Educ 2016;16(1):Article250. PMID:27681933 doi:10.1186/s12909-016-0773-3

167. Mafinejad MK, Arabshahi SKS, Monajemi A, Jalili M, Soltani A, Rasouli J. Use of multi-response format test in the assessment of medical students' critical thinking ability. J Clin Diagn Res 2017;11(9):10-3. PMID:29207742 doi:10.7860/JCDR/2017/24884.10607

168. Puthiaparampil T. Assessment analysis: how it is done. MedEdPublish 2017;6:Article7. doi:10.15694/mep.2017.000142

169. Vander Beken H, Brysbaert M. Studying texts in a second language: the importance of test type. Bil Lang Cog 2017;21(5):1062-74. doi:10.1017/s1366728917000189

170. Lahner FM, Lörwald AC, Bauer D, Nouns ZM, Krebs R, Guttormsen S, et al. Multiple true-false items: A comparison of scoring algorithms. Adv in Health Sci Educ 2018;23(3):455-63. PMID:29189963 doi:10.1007/s10459-017-9805-y

171. Puthiaparampil T, Rahman MM. Very short answer questions: a viable alternative to multiple choice questions. BMC Med Educ 2020;20(1):Article141. PMID:32375739 doi:10.1186/s12909-020-02057-w

172. May MA. Measuring achievement in elementary psychology and in other college subjects. School Soc 1923;17(435):472-6.

173. Remmers HH, Gage NL. Educational Measurement and Evaluation. 2nd ed. New York, NY: Harper & Brothers; 1955.

174. Stanley JC, Hopkins KD. Educational and Psychological Measurement and Evaluation. 5th ed. Englewood Cliffs, NJ: Prentice-Hall; 1972. ISBN:9780132362818

175. Mehrens WA, Lehmann IJ. Measurement and Evaluation in Education and Psychology. 3rd ed. New York, NY: Holt, Rinehart and Winston; 1984. ISBN:9784833701907

176. Ebel RL, Frisbie DA. The Administration and Scoring of Achievement Tests. Essentials of educational measurement. Englewood Cliffs, NJ: Prentice-Hill; 1986.

177. Ebel RL, Frisbie DA. Essentials of Educational Measurement. 5th ed. Englewood Cliffs, NJ: Prentice-Hall; 1991. ISBN:9780132846134

178. Mehrens WA, Lehmann IJ. Measurement and evaluation in education and psychology. 4th ed. New York, NY: Holt, Rinehart and Winston; 1991. ISBN:9780030304071

179. Rogers HJ. Guessing in multiple choice tests. In: Masters GN, Keeves JP, editors. Advances in Measurement in Educational Research and Assessment. Kidlington, UK: Pergamon; 1999.

180. Burton RF. Quantifying the effects of chance in multiple choice and true/false tests: Question selection and guessing of answers. Ass Eval High Educ 2001;26(1):41-50. doi:10.1080/02602930020022273

181. Foster RR, Ruch GM. On corrections for chance in multiple-response tests. J Educ Psychol 1927;18(1):48-51. doi:10.1037/h0070562

182. Little EB. Overcorrection for guessing in multiple-choice test scoring. J Educ Res 1962;55(6):245-52. doi:10.1080/00220671.1962.10882801

183. Holzinger KJ. On scoring multiple response tests. J Educ Psychol 1924;15(7):445-7. doi:10.1037/h0073083

184. Ruch GM, Degraff MH. Corrections for chance and “guess” vs. “do not guess” instructions in multiple-response tests. J Educ Psychol 1926;17(6):368-75. doi:10.1037/h0073222

185. Ruch GM. The objective or new-type examination: An introduction to educational measurement. Chicago, IL: Scott, Doresman and Company; 1929.

186. Kinney LB, Eurich AC. Studies of the true-false examination. Psychol Bull 1933;30(7):505-17. doi:10.1037/h0070031

187. Lincoln EA, Lincoln LL. The preparation of new type testing materials. Testing and the uses of test results. New York, NY: Macmillan; 1935. p. 182-205.

188. Guilford JP. The determination of item difficulty when chance success is a factor. Psychometrika 1936;1(4):259-64. doi:10.1007/BF02287877

189. Votaw DF. The effect of do-not-guess directions upon the validity of true-false or multiple-choice tests. J Educ Psychol 1936;27(9):698-703. doi:10.1037/h0055572

190. Wood HP. Objective test forms for school certificate physics. Br J Educ Psych 1943;13(3):141-6. doi:10.1111/j.2044-8279.1943.tb02733.x

191. Varty JW. Guessing on examinations - is it wortwhile? Educ Forum 1946;10(2):205-12. doi:10.1080/00131724609342257

192. Cronbach LJ. Essentials of Psychological Testing. Harper & Brothers; 1949.

193. Weitzman E, McNamara WJ. Scoring and grading the examination. Constructing classroom examinations: a guide for teachers. 2nd ed. Chicago, IL: Science Research Associates; 1949.

194. Lyerly SB. A note on correcting for chance success in objective tests. Psychometrika 1951;16:21-30. doi:10.1007/BF02313424

195. Coombs CH, Milholland JE, Womer FB. The assessment of partial knowledge. Educ Psychol Meas 1953;16(1):13-37. doi:10.1177/001316445601600102

196. Bradfield JM, Moredock HS. Measurement and Evaluation in Education. New York, NY: Macmillan; 1957.

197. Graesser RF. Guessing on multiple-choice tests. Educ Psychol Meas

1958;18(3):617-20. doi:10.1177/001316445801800316

198. Anastasi A. Psychological Testing. 2nd ed. New York, NY: Macmillan; 1961.

199. Glass GV, Wiley DE. Formula scoring and test reliability. J Educ Meas 1964;1(1):43-9. doi:10.1111/j.1745-3984.1964.tb00150.x

200. Cureton EE. The correction for guessing. J Exp Educ 1966;34(4):44-7. doi:10.1080/00220973.1966.11010953

201. Little EB. Overcorrection and undercorrection in multiple-choice test scoring. J Exp Educ 1966;35(1):44-7. doi:10.1080/00220973.1966.11010968

202. Storey AG. A review of evidence or the case against the true-false item. J Educ Res 1966;59(6):282-5. doi:10.1080/00220671.1966.10883357

203. Lennox B. Marking multiple-choice examinations. Brit J Med Educ 1967;1(3):203-11. PMID:6080736 doi:10.1111/j.1365-2923.1967.tb01698.x

204. Nunnally JC. Psychometric Theory. New York, NY: McGraw-Hill; 1967.

205. Hill GC, Woods GT. Multiple true-false questions. Sch Sci Rev 1969;50(173):919-22.

206. Weitzman RA. Ideal multiple-choice items. J Am Stat Assoc 1970;65(329):71-89. doi:10.2307/2283576

207. Collet LS. Elimination scoring: an empirical evaluation. J Educ Meas 1971;8(3):209-14. doi:10.1111/j.1745-3984.1971.tb00927.x

208. Thorndike RL. Educational Measurement. 2nd ed. Washington, DC: American Council on Education; 1971. ISBN:9780826812711

209. Oosterhof AC, Glasnapp DR. Comparative reliabilities and difficulties of the multiple-choice and true-false formats. J Exp Educ 1974;42(3):62-4. doi:10.1080/00220973.1974.11011479

210. Quereshi MY. Performance on multiple-choice tests and penalty for guessing. J Exp Educ 1974;42(3):74-7. doi:10.1080/00220973.1974.11011481

211. Choppin B. Guessing the answer on objective tests. Br J Educ Psychol 1975;45(2):206-13. doi:10.1111/j.2044-8279.1975.tb03245.x

212. Robbins E. Completion and true/false items. Nurs Times 1975;71(44):1751-2. PMID:1196953

213. Frary RB, Cross LH, Lowry SR. Random guessing, correction for guessing, and reliability of multiple-choice test scores. J Exp Educ 1977;46(1):11-5. doi:10.1080/00220973.1977.11011603

214. Benson J, Crocker L. The effects of item format and reading ability on objective test performance: A quastion of validity. Educ Psychol Meas 1979;39(2):381-7. doi:10.1177/001316447903900217

215. Koeslag JH, Melzer CW, Schach SR. Inversion in true/false and in multiple choice questions - a new form of item analysis. Med Educ 1979;13(6):420-4. PMID:537531 doi:10.1111/j.1365-2923.1979.tb01201.x

216. Bergman J. Understanding Educational Measurement and Evaluation. Boston, MA: Houghton Mifflin; 1981. ISBN:9780395307823

217. Koeslag JH, Melzer CW, Schach SR. Penalties in multiple-choice and true-false questions. S Afr Med J 1983;63(1):20-2. PMID:6849146

218. Grosse ME, Wright BD. Validity and reliability of true-false tests. Educ Psychol Meas 1985;45(1):1-13. doi:10.1177/0013164485451001

219. Ellington H. Objective Questions. Teaching and Learning in Higher Education, 21. Aberdeen: Scottish Central Institutions Committee for Educational Development; 1987.

220. Sax G. Principles of Educational and Psychological Measurement and Evaluation. 3rd ed. Belmont, CA: Wadsworth; 1989. ISBN:9780534099787

221. Gronlund NE, Linn RL. Measurement and evaluation in teaching. 6th ed. New York, NY: Macmillan; 1990. ISBN:9780023481116

222. Ory JC, Ryan KE. Tips for improving testing and grading. Newbury Park, CA: Sage Publications Inc.; 1993. ISBN:9780803949737

223. Nunnally JC, Bernstein IH. Psychometric Theory. 3rd ed. New York, NY: McGraw-Hill; 1994. ISBN:9780070478497

224. Beullens J, Jaspaert H. Het examen met meerkeuzevragen [Multiple choice examination]. Ned Tijdschr Geneeskd 1999;55(7):529-35.

225. Oosterhof A. Classroom Applications of Educational Measurement. 3rd ed. Upper Saddle River, NJ: Prentice-Hall; 2001. ISBN:9780135203880

226. Petz B. Penalizirati ili ne penalizirati pogrešne odgovore u testovima znanja alternativnog tipa [To penalize or not to penalize false answers in the achievement tests of the alternative type]. Revija za Psihologiju 1978;8(1-2):49-56.

227. Slakter MJ. The effect of guessing strategy on objective test scores. J Educ Meas 1968;5(3):217-21. doi:10.1111/j.1745-3984.1968.tb00629.x

228. Bush M. A multiple choice test that rewards partial knowledge. Journal of Further and Higher Education 2001;25(2):157-63. doi:10.1080/03098770120050828

229. Gupta RK, Penfold DME. Correction for guessing in true-false tests: An experimental approach. Brit J Educ Psychol 1961;31(P3):249-56. doi:10.1111/j.2044-8279.1961.tb01714.x

230. Asker WM. The reliability of tests requiring alternative responses. J Educ Res 1924;9(3):234-40. doi:10.1080/00220671.1924.10879451

231. Gupta RK. A new approach to correction in true false tests. Educ Psychol (Delhi) 1957;4(2):63-75.

232. Sanderson PH. The ‘don’t know’ option in MCQ examinations. Br J Med Educ 1973;7(1):25-9. PMID:4723448 doi:10.1111/j.1365-2923.1973.tb02206.x

233. Anderson J. Marking of multiple choice questions. The multiple choice question in

medicine. 2nd ed. London, UK: Pitman Books Limited; 1982. p. 45-58.

234. Gajjar S, Sharma R, Kumar P, Rana M. Item and test analysis to identify quality Multiple Choice Questions (MCQs) from an assessment of medical students of Ahmedabad, Gujarat. Indian J Community Med 2014;39(1):17-20. PMID:24696535 doi:10.4103/0970-0218.126347

235. Kohs SC. High test scores attained by subaverage minds. Psychol Bull 1920;17(1):1-5. doi:10.1037/h0064475

236. Chapman JC. Individual injustice and guessing in the true-false examination. J Appl Psychol 1922;6(4):342-8. doi:10.1037/h0076011

237. Hahn HH. A criticism of tests requiring alternative responses. J Educ Res 1922;6(3):236-41. doi:10.1080/00220671.1922.10879299

238. McCall WA. How to Measure in Education. New York, NY: Macmillan; 1922.

239. West PV. A critical study of the right minus wrong method. J Educ Res 1923;8(1):1-9. doi:10.1080/00220671.1923.10879376

240. Batson WH. Reliability of the true-false form of examination. Educ Admin Supervision 1924;10:95-102.

241. Miller GF. Tinkering with a true-false test. Proc Okla Acad Sci 1925;5:25-30.

242. Weidemann CC. How to construct the true-false examination. New York, NY: Teachers’ College, Columbia University; 1926.

243. Palmer I. New type examinations in physical education. Am Physical Educ Rev 1929;34(3):151-6. doi:10.1080/23267224.1929.10652100

244. Jensen MB. An evaluation of three methods of presenting True-False examinations: visual, oral and visual-oral. School Soc 1930;32(829):675-7.

245. Barton WA. Improving the true-false examination. School Soc 1931;34(877):544-6.

246. Granich L. A technique for experimentation on guessing on objective tests. J Educ

Psychol 1931;22(2):145-56. doi:10.1037/h0072728

247. Peters CC, Martz HB. A study of the validity of various types of examinations. School Soc 1931;33(845):336-8.

248. Krueger WCF. An experimental study of certain phases of a true-false test. J Educ Psychol 1932;23(2):81-91. doi:10.1037/h0073943

249. Lee JM, Symonds PM. New-type or objective tests: A summary of recent investigations. J Educ Psychol 1933;24(1):21-38. doi:10.1037/h0072226

250. Soderquist HO. A new method of weighting scores in a true-false test. J Educ Res 1936;30(4):290-2. doi:10.1080/00220671.1936.10880670

251. Moore CC. Factors of chance in the true-false examination. J Genet Psychol 1938;53(1):215-29. doi:10.1080/08856559.1938.10533806

252. Swineford F. The measurement of a personality trait. J Educ Psychol 1938;29(4):295-300. doi:10.1037/h0058735

253. Etoxinod S. How to checkmate certain vicious consequences of true-false tests. Etoxin 1940;61:223-7.

254. Moore CC. The rights-minus-wrongs method of correcting chance factors in the true-false examination. J Genet Psychol 1940;57(2):317-26. doi:10.1080/08856559.1940.10534539

255. Cronbach LJ. An experimental comparison of the multiple true-false and multiple multiple-choice tests. J Educ Psychol 1941;32(7):533-43. doi:10.1037/h0058518

256. Weidemann CC. The “omission” as a specific determiner in the true-false examination. J Educ Psychol 1941;32(7):435-9. doi:10.1037/h0074950

257. Cruze WW. Measuring the results of learning. Educational Psychology. New York, NY: The Ronald Press Company; 1942. p. 343-80.

258. Gilmour WA, Gray DE. Guessing on true-false tests. Educ Res Bull 1942;21(1):9-12.

259. Cronbach LJ. Studies of acquiescence as a factor in the true-false test. J Educ Psychol 1942;33(6):401-15. doi:10.1037/h0054677

260. Mead AR, Smith BM. Does the true-false scoring formula work? Some data on an old subject. J Educ Res 1957;51(1):47-53. doi:10.1080/00220671.1957.10882437

261. Harden RM, Gleeson FA. Assessment of clinical competence using an objective structured clinical examination (OSCE). Med Educ 1979;13(1):39-54. PMID:763183 doi:10.1111/j.1365-2923.1979.tb00918.x

262. Fleming PR. The profitability of ‘guessing’ in multiple choice question papers. Med Educ 1988;22(6):509-13. PMID:3226344 doi:10.1111/j.1365-2923.1988.tb00795.x

263. Jacobs LC, Chase CI. Developing and using tests effectively. San Francisco, CA: Jossey-Bass Inc.; 1992. ISBN:9781555424817

264. Hammond EJ, McIndoe AK, Sansome AJ, Spargo PM. Multiple-choice examinations: adopting an evidence-based approach to exam technique. Anaesthesia 1998;53(11):1105-8. PMID:10023280 doi:10.1046/j.1365-2044.1998.00583.x

265. Chase CI. Contemporary Assessment for Educators. New York, NY: Longman; 1999. ISBN:9780801313721

266. Downing SM. The effects of violating standard item writing principles on tests and students: the consequences of using flawed test items on achievement examinations in medical education. Adv Health Sci Educ Theory Pract 2005;10(2):133-43. PMID:16078098 doi:10.1007/s10459-004-4019-5

267. Dijksterhuis MG, Scheele F, Schuwirth LW, Essed GG, Nijhuis JG, Braat DD. Progress testing in postgraduate medical education. Med Teach 2009;31(10):e464-e8. PMID:19877854 doi:10.3109/01421590902849545

268. Staffelbach EH. Weighting responses in True-False examinations. J Educ Psychol 1930;21(2):136-9. doi:10.1037/h0072266

269. Gates AI. The true-false test as a measure of achievement in college courses. J Educ Psychol 1921;12(5):276-87. doi:10.1037/h0074436

270. Rao NJ. A note on the evaluation of the true-false and similar tests of the new-type examination. Indian J Psychol 1937;12:176-9.

271. Kirstges T. Gerechte Noten: Zur Gestaltung von Notensystemen für die Beurteilung von Leistungen in Klausuren [Fair grades: designing grading systems for assessing performance in exams]. Neue Hochschule 2007;48(3):26-31.

272. Frary RB. NCME instructional module: formula scoring of multiple-choice tests (correction for guessing). Educ Meas 1988;7(2):33-8. doi:10.1111/j.1745-3992.1988.tb00434.x

273. Lukas J, Melzer A, Much S. Auswertung von Klausuren im Antwort-Wahl-Format [Evaluation of multiple-choice examinations]. Halle (Saale): Center for media-enhanced Learning and Teaching (LZZ) of the Martin Luther University of Halle-Wittenberg; 2017. ISBN:9783868298727

274. Tarrant M, Knierim A, Hayes SK, Ware J. The frequency of item writing flaws in multiple-choice questions used in high stakes nursing assessments. Nurse Education Today 2006;26(8):662-71. PMID:17014932 doi:10.1016/j.nedt.2006.07.006

275. de Laffolie J, Visser D, Hirschburger M, Turial S. „Cues“ und „pseudocues“ in chirurgischen MC-Fragen des deutschen Staatsexamens [Cues and pseudocues in surgical multiple choice questions from the German state examination]. Der Chirurg 2017;88(3):239-43. PMID:27678403 doi:10.1007/s00104-016-0291-1

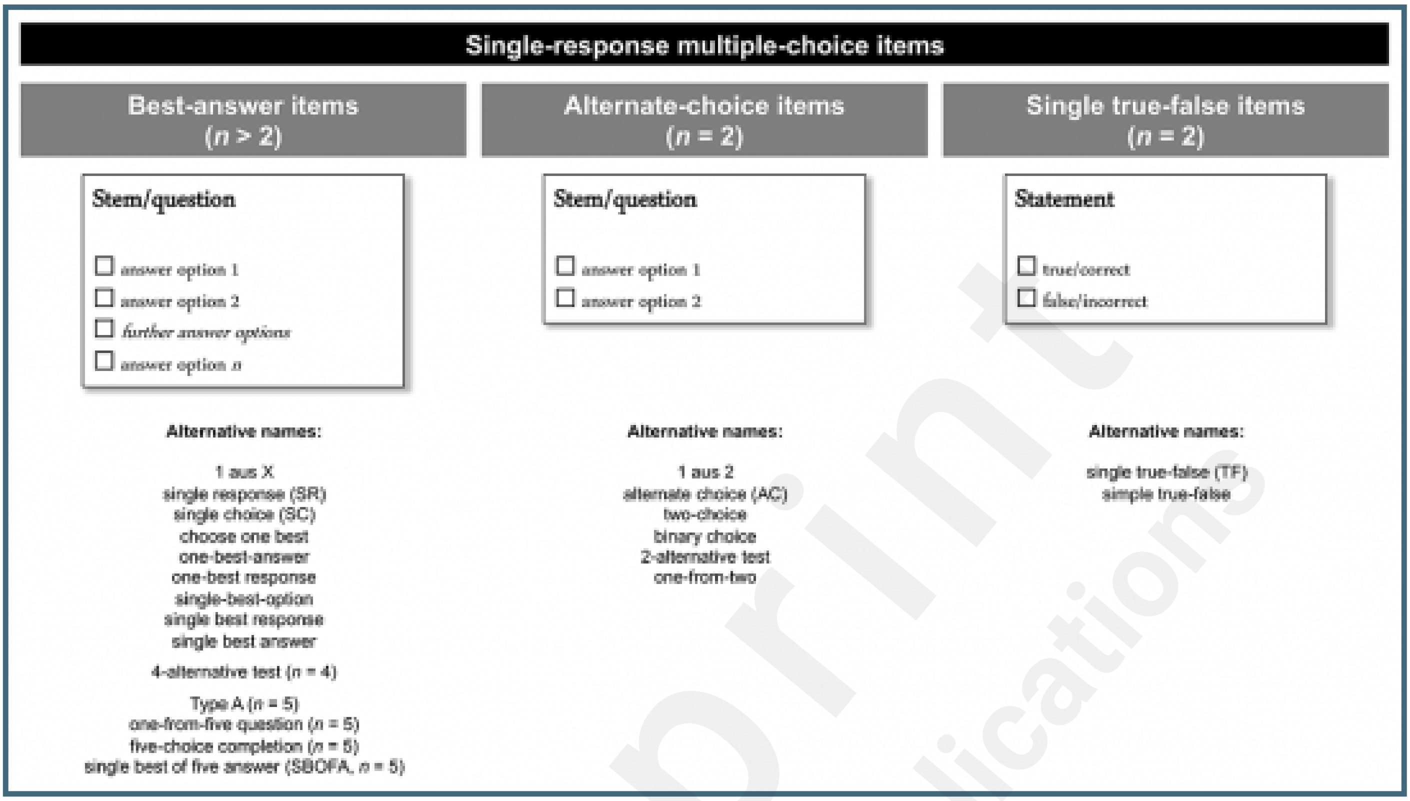
276. Kanzow P, Schmidt D, Herrmann M, Wassmann T, Wiegand A, Raupach T. Use of multiple-select multiple-choice items in a dental undergraduate curriculum: retrospective application of different scoring methods. JMIR Med Educ 2023;9:e43792. PMID:36841970 doi:10.2196/43792

277. Kubinger KD. Gutachten zur Erstellung „gerichtsfester” Multiple-Choice-Prüfungsaufgaben [Expert opinion on the creation of “lawful” multiple-choice items]. Psychol Rundschau 2014;65(3):169-78. doi:10.1026/0033-3042/a000218

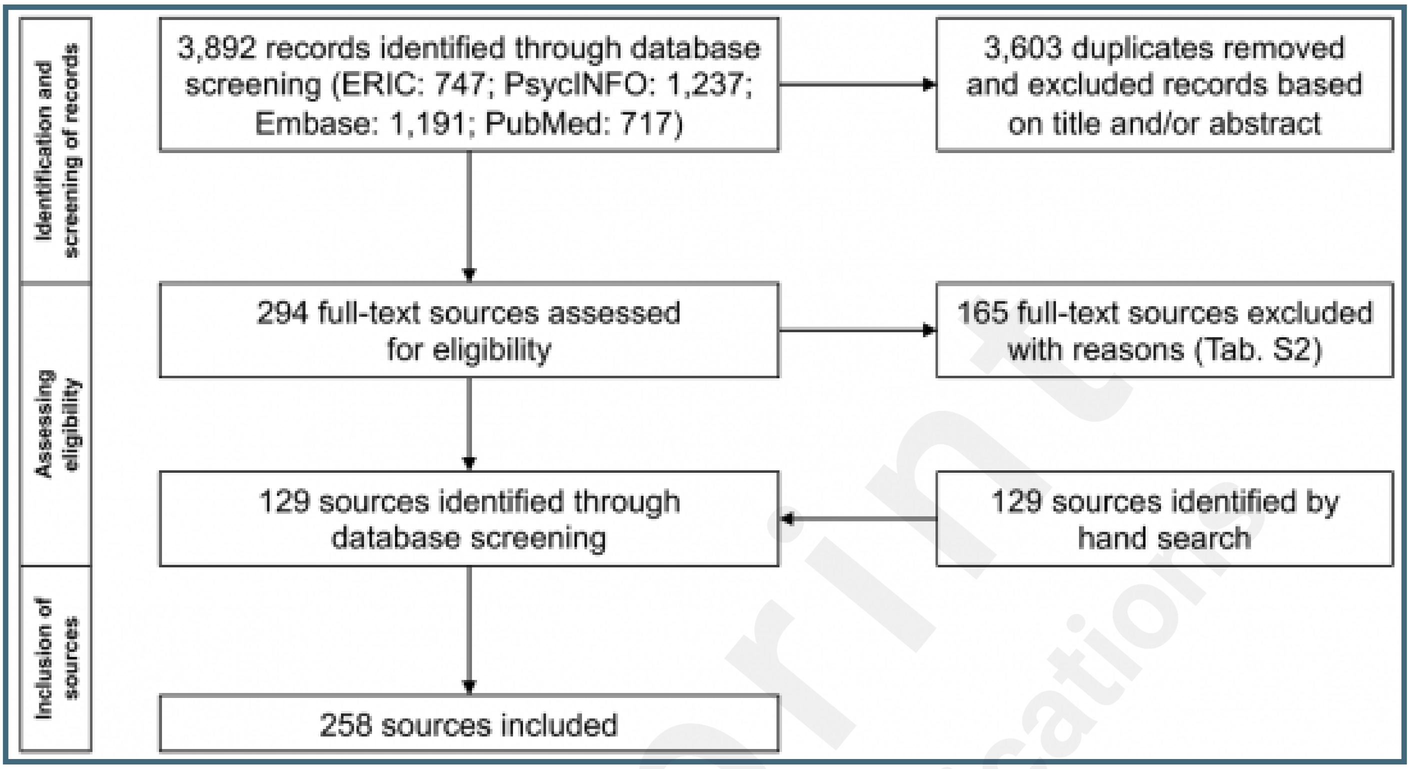
Supplementary Files

Figures

Examples of three different multiple-choice items in single-choice format and alternative designations used in the literature (no claim to completeness).

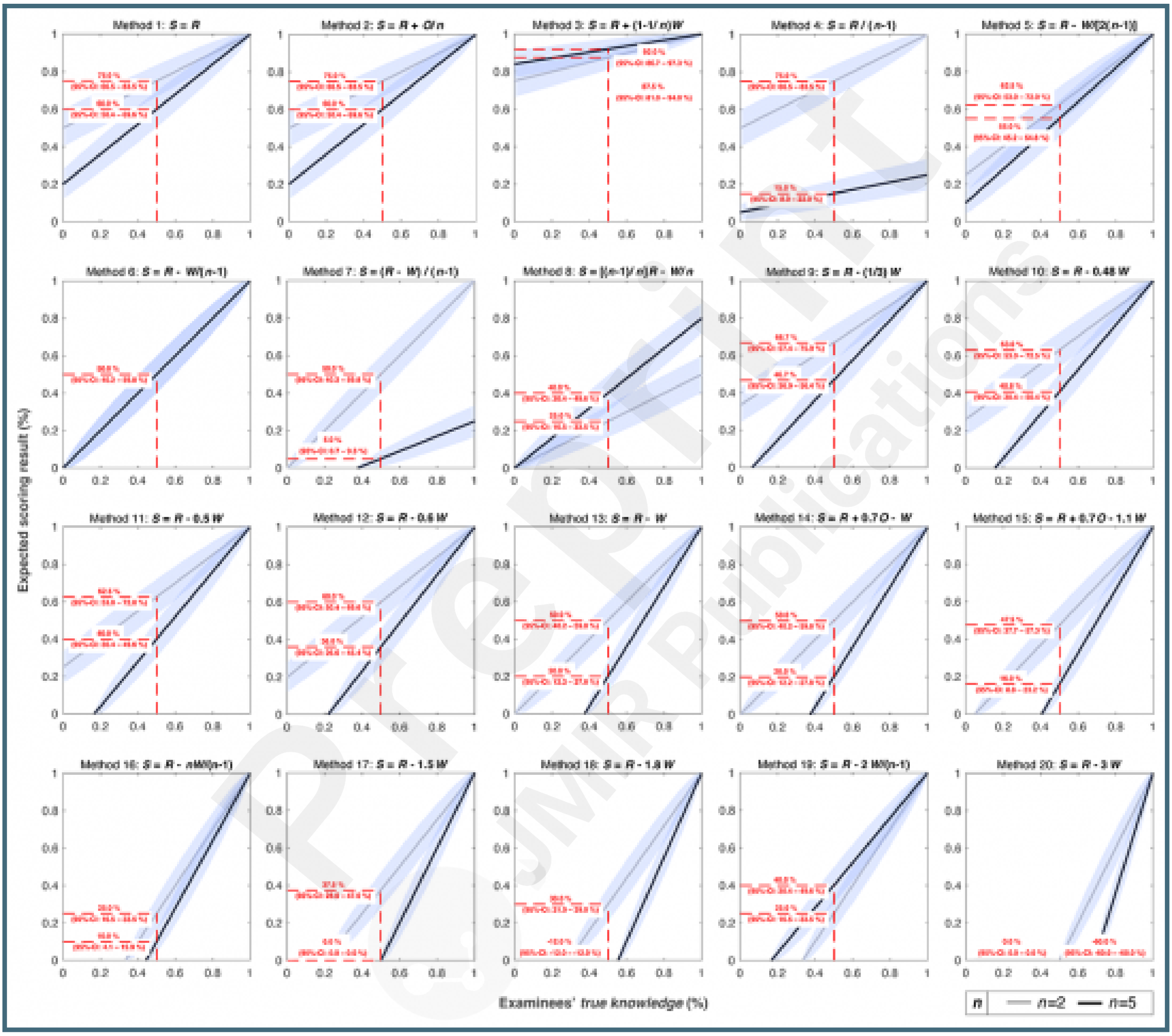


Flow diagram of systematic literature search.


Identification and screening of records
Assessing eligibility
Inclusion of sources
3,892 records identified through database screening (ERIC: 747; PsycINFO: 1,237; Embase: 1,191; PubMed: 717)
3,603 duplicates removed and excluded records based on title and/or abstract
294 full-text sources assessed for eligibility
165 full-text sources excluded with reasons (Tab. S2)
129 sources identified through database screening
129 sources identified by hand search
258 sources included

Equation for the calculation of the expected scoring result (f = credit points awarded for a correctly marked item [$i = 1$] or an incorrectly marked item [$i = 0$] depending on the scoring method used; k = examinees' *true knowledge* [0 ? k ? 1]; n = number of answer options per item; $x = 1$ if the correct answer option is selected by *true knowledge*, otherwise $x = 0$; in the equation shown, 0^0 is defined as 1).

$$Expected\ scoring\ result = \sum_{i=0}^{1}\sum_{x=0}^{i}\left(k^{x} * (1-k)^{1-x}\right) * \frac{\binom{n-1}{1-i}}{\binom{n-x}{1-x}} * f_i$$


Method 1: $S = R$
Method 2: $S = R + O/n$
Method 3: $S = R + (1-1/n)W$
Method 4: $S = R/(n-1)$
Method 5: $S = R - W/[2(n-1)]$
Method 6: $S = R - W/(n-1)$
Method 7: $S = (R - W)/(n-1)$
Method 8: $S = [(n-1)/n]R - W/n$
Method 9: $S = R - (1/3)W$
Method 10: $S = R - 0.48W$
Method 11: $S = R - 0.5W$
Method 12: $S = R - 0.6W$
Method 13: $S = R - W$
Method 14: $S = R + 0.7O - W$
Method 15: $S = R + 0.7O - 1.1W$
Method 16: $S = R - nW/(n-1)$
Method 17: $S = R - 1.5W$
Method 18: $S = R - 1.8W$
Method 19: $S = R - 2W/(n-1)$
Method 20: $S = R - 3W$
Expected scoring result (%)
Examinees' true knowledge (%)
0
0.2
0.4
0.6
0.8
1
n n=2 n=5

Multimedia Appendixes

Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews Checklist.
URL: http://asset.jmir.pub/assets/7df1a80ad6e828ee5b93cf13f693aa39.docx

Excluded sources after screening of full-texts.
URL: http://asset.jmir.pub/assets/c019b992a23c77716cd0d3d511620638.docx

High resolution version of Figure 4.
URL: http://asset.jmir.pub/assets/1648c6eb1f4e8cee936890404194f258.pdf

CONSORT (or other) checklists

Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews Checklist.
URL: http://asset.jmir.pub/assets/ddddb073fcc167a2ed1059fd046981a2.pdf

TOC/Feature image for homepages

Examinee marking a single-response multiple-choice examination.