Adv in Health Sci Educ (2011) 16:211–221
DOI 10.1007/s10459-010-9256-1

Pick-N multiple choice-exams: a comparison of scoring algorithms

Daniel Bauer · Matthias Holzer · Veronika Kopp · Martin R. Fischer

Received: 16 August 2010 / Accepted: 20 October 2010 / Published online: 31 October 2010

Abstract To compare different scoring algorithms for Pick-N multiple correct answer multiple-choice (MC) exams regarding test reliability, student performance, total item discrimination and item difficulty. Data from six 3rd year medical students' end of term exams in internal medicine from 2005 to 2008 at Munich University were analysed (1,255 students, 180 Pick-N items in total). Scoring Algorithms: Each question scored a maximum of one point. We compared: (a) Dichotomous scoring (DS): One point if all true and no wrong answers were chosen. (b) Partial credit algorithm 1 (PS_{50}): One point for 100% true answers; 0.5 points for 50% or more true answers; zero points for less than 50% true answers. No point deduction for wrong choices. (c) Partial credit algorithm 2 ($PS_{1/m}$): A fraction of one point depending on the total number of true answers was given for each correct answer identified. No point deduction for wrong choices. Application of partial crediting resulted in psychometric results superior to dichotomous scoring (DS). Algorithms examined resulted in similar psychometric data with PS_{50} only slightly exceeding $PS_{1/m}$ in higher coefficients of reliability. The Pick-N MC format and its scoring using the PS_{50} and $PS_{1/m}$ algorithms are suited for undergraduate medical examinations. Partial knowledge should be awarded in Pick-N MC exams.

Keywords Undergraduate assessment · Multiple choice · Scoring · Test psychometrics · Pick-N · Undergraduate medical education

D. Bauer (✉) · V. Kopp · M. R. Fischer
Faculty of Health, Institute for Teaching and Educational Research in Health Sciences, Witten/Herdecke University, Alfred-Herrhausen-Strasse 50, 58448 Witten, Germany
e-mail: daniel.bauer@uni-wh.de

M. Holzer · M. R. Fischer
Medizinische Klinik—Innenstadt, Medical Education Unit, Munich University Hospital, Ziemssenstrasse 1, 80336 Munich, Germany

Introduction

Background

Commonly used in the context of summative assessment in undergraduate medical education, Multiple Choice Questions (MCQs) need little introduction. Like any other means of assessment they have advantages (high levels of standardization, transparent marking, economic testing of large classes, and broad sampling of knowledge) (Norcini et al. 1984; Schuwirth and van der Vleuten 2003), and disadvantages, often so when poorly written (testing knowledge of isolated facts or just improperly phrased) and used exclusively without being supplemented by other means of assessment for the purpose of triangulation. This alleged limitation of MCQs to testing little more than isolated, factual knowledge has begun to crumble recently. The literature shows that well composed MCQs can test higher order diagnostic thinking and application of knowledge, evaluating the examinee's ability to integrate, synthesize and judge medical information (Norcini et al. 1984; Case and Swanson 2001; Coderre et al. 2004). However, cueing effects through pre-worded answers and answer hints through inexpert phrasing are an important issue (Haladyna et al. 2002).

MCQs continue to be an important part of the assessment toolbox, but the format generally used is the single best answer question (generally one-out-of-five, type A). But is this the most suited question format for clinical problems, which might have more than just one correct solution?

As the curriculum and learning progress, students reach a point where hypotheses have to be generated and refined, and diagnostic and therapeutic options weighed (Kassirer and Kopelman 1991). So, in 1996 Ripkey et al. added a new item format to the MC typology that had the benefits of MCQs but could present "reasonably realistic clinical challenges", i.e. which takes into account that clinical problems can have more than just one solution: PICK-N, a multiple best answer format described as a mixture of Multiple True/False, Patient Management Problem, Key Feature Problem and Extended Matching Question (Case and Swanson 2001; Ripkey et al. 1996). Pick-N items prompt examinees to pick several correct out of various answer possibilities (n out of many).[1]

Thereby Pick-N items also act as an alternative to the no longer used K-type (1 out of five combinations of answer options) (Albanese 1993) and permit a relatively easy re-write of negatively worded items (Case and Swanson 2001; Haladyna and Downing 1989a), and permit rewarding partial knowledge (Lord 1963; Ben-Simon et al. 1997). However, most research on item writing has been done on less complicated case vignettes and with less advanced students.

In contrast to single best choice items however, for which a dichotomous grading system is obviously a good choice, there are more choices about how to grade Pick-N items.

Scoring in written assessment of clinical reasoning

The open-ended question (e.g. Short Answer Question (SAQ) or short essay), the mother of assessment in clinical reasoning, minimizes cueing to some extent and offers ample flexibility but takes more time than MCQs in generation, answering and grading (Schuwirth and van der Vleuten 2003; Epstein 2007). Detailed directions for weighing the possible answers are crucial when utilizing open-ended questions and examiner variability

[1] See "Appendix" for example.

in marking has to be considered (Wakeford and Roberts 1984; Nendaz and Tekian 1999). Few recommendations however exist on grading multiple best answer items, even for the closer relatives of Pick-N:

Albanese evaluated multiple response credit algorithms for the Multiple True False format (MTF, or Type-X: multiple statements that are either correct or incorrect, in this case four true/false options per item) and found partial crediting to be superior to dichotomous scoring regarding test reliability and some indicators of validity (Albanese and Sabers 1988). These findings were later confirmed by Itten in 1997 in a similar approach (Itten and Krebs 1997).

Key feature (KF) problems focus mainly on clinical decision making and are defined more by content than by format. They have the highest added value when in electronic exams where Long Menu answer prompts can be used as quasi open-ended questions (Schuwirth and van der Vleuten 2003; Rotthoff et al. 2006). KF problems assessed with Short Menu or Multiple Choice formats are normally scored dichotomously.

In Script Concordance Testing (SCT), weighing of answer alternatives is done by a panel of content experts, whose result is usually not one single best answer but any of the Likert style answers can get partial credit, depending on how many members of the panel regarded it as the best answer (Fournier et al. 2008). Bland et al. however suggest that the psychometrics in this kind of aggregate scoring are quite similar to single best answer scoring (Bland et al. 2005).

Extended Matching Questions (EMQs, also referred to as R-type MCQs) are yet another single best answer MCQ variant where the examinee chooses his answer from a list of options (one out of many, up to 26). Generally two or more different questions refer to that answer list. EMQs combine elements of SAQs and traditional one-out-of-five MCQs. The literature suggests this format might be well suited for the assessment of clinical reasoning (Case and Swanson 1993; Beullens et al. 2005; Beullens et al. 2006), whereas long answer lists might be adding little to discriminatory power of an item while taking up additional testing time (Swanson et al. 2008).

Concerning the grading of Pick-N items, Ripkey suggested a simple partial crediting algorithm. He compared the psychometrics of 54 Pick-N items with 267 classic A-type MCQs in the context of the USMLE Step 2, favouring a grading system that awards 1/m points (m representing the number of correct answers) per item, pointing out benefits regarding item discrimination and difficulty (Ripkey et al. 1996). Items with more answers chosen than allowed were awarded zero points. Apparently no points were deducted for wrong answers. However, Ripkey identifies several limitations to the study carried out, for instance item authors and examinees were not well used to the format and only 54 Pick-N items were included in the study. Conducted in an USMLE context, the implications for undergraduate assessment were not clear and only one partial credit algorithm (1/m) examined. These results so far have not been reproduced and little is known about the comparative psychometric characteristics of different partial scoring algorithms for the Pick-N question format in an undergraduate setting.

Scoring Pick-N multiple best answer questions

This study aims to confirm Ripkey's findings and to identify an optimal partial credit scoring algorithm for Pick-N (multiple best select) multiple choice items with regard to psychometrics (test reliability, item discrimination, item difficulty) by comparing a dichotomous and two partial credit scoring algorithms in an undergraduate assessment setting.

Methods

Empirical data from six consecutive, independently scored end-of-term exams in internal medicine (endocrinology, rheumatology/nephrology, infectious diseases/pneumology, gastroenterology/hepatology, hemato-oncology, and cardiology/angiology) between summer 2005 and winter 2008 from the Ludwig-Maximilians-University Medical Faculty in Munich, Germany, were included in this study. Students represented whole classes in their 3 rd year with a mean of 209 students per exam, totalling 1,255 examinees.

Exams consisted of 60 items, incorporating Pick-N (M = 28.0, SD = 3.3, total = 175) and A-type (M = 31.2, SD = 3.0, total 180) items. In total five items were excluded after the exams for formal flaws identified in post-exam review (M = 0.83, SD = 0.08). Time on task was 90 s per item, or 90 min for each exam.

As this study analyses Pick-N items characteristics, Pick-N subsets of the evaluated exams are treated as standalone tests. In a few cases, psychometric data of A-type subsets is given for reasons of quality control and comparability.

Items were designed by tutors and lecturers directly involved in teaching and are consistent with the faculty's learning objectives for internal medicine and its disciplines published and reviewed by a central team of content experts as well as assessment experts. The latter had at least 1 year of experience with the Pick-N format and followed generally accepted item writing guidelines in form and content (Case and Swanson 2001; Haladyna et al. 2002; Haladyna and Downing 1989b; Krebs 2004). Flawed items were improved after consultation with original item designers or discarded. No German National Board of Medical Examiners (IMPP) items were used.

Each correctly solved item could be awarded a maximum of one point with no negative marking rule in action (no point deduction for wrong answers). The number of corrects answers to each question was given. If a student marked more answers as correct than they were asked to for a given question, the number of extra marks was subtracted from the number of correctly identified answers to that question, the rest then counted towards the score.

With these common rules the following item scoring algorithms were compared:

1. Dichotomous scoring ("DS"): One point if all correct and no wrong answer choices were identified. No partial credit is awarded. The strictest of algorithms examined, DS follows the rationale that patient treatment cannot tolerate mistakes and exams have to test accordingly.
2. Partial credit scoring 1 ("PS_{50}"): One point if all correct and no wrong answers marked. Half a point awarded if not all but at least 50% of correct answers chosen; else zero points. PS_{50} values partial knowledge, but only after a certain threshold is passed (in this case 50%). A hybrid between DS and $PS_{1/m}$, on the grounds that a fraction of knowledge is not enough to treat a patient adequately while a certain portion of knowledge will allow adequate but possibly suboptimal treatment.
3. Partial credit scoring 2 ("$PS_{1/m}$"): 1/m points awarded for each of m correct answers possible. Most lenient of algorithms, $PS_{1/m}$ awards partial credit for any correct choice made—as any proper choice made in patient treatment might be beneficial for the patient.

An item with n = four correct answers would then be scored as follows:

m	0	1	2	3	4
DS	0.00	0.00	0.00	0.00	1.00
P_{50}	0.00	0.00	0.50	0.50	1.00
$P_{1/m}$	0.00	0.25	0.50	0.75	1.00

Points earned in Pick-N item with n = 4 correct answers. m denominates the number of correct answers chosen by candidate

To answer an item, all answer possibilities identified as correct had to be marked on a machine-readable form. The number of correct answers (ranging from 2 to 6) was given at the end of each question stem. The number of distractors ranged from 3 to 10, resulting in a list of 5–15 choices. Each answer list comprised at least as many distractors as correct answers.

Data from answer sheets was psychometrically analysed using the aforementioned algorithms.

To approximate internal test consistency (i.e. test reliability), Cronbach's alpha was calculated for each combination of exam and algorithm (Downing 2004). These empirical data were then projected to 100 item tests by means of the Spearman-Brown prediction formula to demonstrate more easily comparable differences and similarities in test reliability.

Further psychometric criteria of interest were item difficulty and total item correlations:

"Item difficulty p" was defined as the mean of points earned for the particular item. A low rate of items with difficulty levels below 0.4 ("item rather too difficult") or above 0.8 ("item rather easy") was considered a quality feature (Möltner et al. 2006). The threshold of 0.8 might seem arbitrary and the discussion on the optimal distribution of p values is ongoing: Items with p higher than about 0.8 (meaning most of the examinees got the right answers) tend to add little extra information for differentiating between high and low performers and can seem uneconomical. On the other hand even if an item is easy to answer, it might have been included in the test to cover an important topic of the test blueprint and can thus add to the validity of the test.

Item total correlations were calculated as unadjusted[2] Pearson-Bravais coefficient r, r indicating an item's power of representing the overall test success. Values above 0.2 were estimated a trait of quality (Möltner et al. 2006). Setting the limit for r at 0.2 is, similar to p, a matter of policy. The closer p gets to 1, the more r will probably decrease, since the discriminatory power of an item that every examinee got right is zero (the same applies to r in extremely difficult items). Speaking in terms of economic testing, items with $r < 0.2$ are a bit of a waste of resources, while there are good reasons to include such items for test validity considerations.

Items with unsatisfactory psychometrics were not excluded from this study so as to retain realistic conditions and include the whole bandwidth of good and suboptimal items into the calculations. In a few cases post examination reviews and student evaluation revealed faulty items that had nonetheless passed pre-test reviews, which were excluded from this study (max. 2 per test).

The Faculty Ethics Committee declared no vote was necessary for this study. Data on individuals were treated with proper discretion.

[2] i.e., containing the autocorrelation.

Results

Test reliability

Test reliability as approximated by Cronbach's alpha was lowest for dichotomous scoring ($\alpha_{min} = 0.59$, $\alpha_{max} = 0.74$, $M = 0.65$, $SD = 0.06$) except in the winter term exam 05/06. Application of PS_{50} resulted in highest test reliabilities ($\alpha_{min} = 0.608$; $\alpha_{max} = 0.780$, $M = 0.681$, $SD = 0.05$), the values for $PS_{1/m}$ minimally behind ($\alpha_{min} = 0.609$; $\alpha_{max} = 0.745$, $M = 0.672$ ($SD = 0.05$).

Including the A-type items into the calculation of Cronbach's alpha, test reliabilities were between 0.70 and 0.91; see Table 1 for details.

Test reliabilities were then projected to 100 item tests using Spearman-Brown prediction formula. Reliabilities then differ between 0 and 0.03 between dichotomous and partial scoring with little or no difference at all between the different partial scoring algorithms. The number of items necessary to achieve an $\alpha \geq 0.8$ is usually several items higher with dichotomous scoring than with partial credit, see Table 2 for details.

Item difficulty p

Mean item difficulties calculated dichotomously (DS: $p_{min} = 0.52$, $p_{max} = 0.65$, $M(p) = 0.60$) were consistently below those calculated through partial crediting ($PS_{1/m}$: $p_{min} = 0.79$, $p_{max} = 0.84$, $M(p) = 0.82$) in full concordance with previous findings (Ripkey et al. 1996). Applying PS_{50} resulted in easier to answer items (PS_{50}: $p_{min} = 0.73$, $p_{max} = 0.81$, $M(p) = 0.78$); though the mean was still below that of $PS_{1/m}$, see Table 3 for details.

When dichotomously scored about 20% of Pick-N items were regularly rather too difficult, whereas this held true for only one single item (0.03%) when awarding partial credits. Accordingly, items tended to be easier when applying PS_{50} or $PS_{1/m}$, where the upper threshold of 0.8 was regularly exceeded ($PS_{1/m}$: 51% of items, PS_{50} 42%).

Table 1 Test reliabilities as approximated by Cronbach's alpha of Pick-N and A-type subsets as well as for the complete tests (applying PS_{50} for scoring Pick-N)

Term	Pick-N				A-type		Pick-N + A-type	
		DS	PS_{50}	$PS_{1/m}$			PS_{50}	
	n (items)	Alpha			n (items)	Alpha	n (items)	Alpha
Summer 05	30	0.59	0.65	0.65	30	0.64	60	0.77
Winter 05/06	30	0.61	0.61	0.61	30	0.50	60	0.70
Summer 06	31	0.65	0.68	0.67	27	0.72	58	0.78
Winter 06/07	28	0.74	0.76	0.75	32	0.84	60	0.87
Summer 07	27	0.68	0.71	0.69	32	0.71	59	0.91
Winter 07/08	22	0.60	0.68	0.67	36	0.64	58	0.78
Mean	28	0.65	0.68	0.67	31	0.68	59	0.80

Table 2 Test reliabilities of the Pick-N subsets projected to 100 item tests through Spearman-Brown projection formula

Term	DS		PS_{50}		$PS_{1/m}$	
	$n_{0,8}$	α_{100}	$n_{0,8}$	α_{100}	$n_{0,8}$	α_{100}
Summer 05	84	0.83	65	0.86	66	0.86
Winter 05/06	77	0.84	78	0.84	78	0.84
Summer 06	66	0.86	58	0.87	61	0.87
Winter 06/07	39	0.91	36	0.92	39	0.91
Summer 07	52	0.89	44	0.90	50	0.89
Winter 07/08	58	0.87	43	0.90	43	0.90
Mean	63	0.87	54	0.88	56	0.88

$n_{0.8}$ denominates the number of items theoretically necessary for a test reliability of $\alpha = 0.8$; α_{100} shows Cronbach's alpha for each exam if it had 100 items

Table 3 Mean item difficulties p of Pick-N items according to algorithm applied

Term		M(p) (SD)			$p < 0.4$			$p > 0.8$		
	n (items)	DS	PS_{50}	$PS_{1/m}$	DS	PS_{50}	$PS_{1/m}$	DS	PS_{50}	$PS_{1/m}$
Summer 05	30	0.65 (0.26)	0.81 (0.14)	0.84 (0.12)	7	0	0	12	15	16
Winter 05/06	30	0.52 (0.28)	0.73 (0.17)	0.79 (0.12)	10	1	1	3	12	14
Summer 06	31	0.64 (0.21)	0.80 (0.11)	0.84 (0.10)	7	0	0	3	16	21
Winter 06/07	28	0.62 (0.22)	0.78 (0.14)	0.82 (0.13)	5	0	0	5	15	19
Summer 07	27	0.65 (0.22)	0.81 (0.12)	0.84 (0.11)	4	0	0	7	14	16
Winter 07/08	22	0.55 (0.26)	0.75 (0.16)	0.80 (0.12)	6	0	0	5	10	14
Mean	28	0.60 (0.24)	0.78 (0.14)	0.82 (0.12)	6.5	0.2	0.2	5.8	13.7	16.7

Item difficulty as mean points achieved per item. Also shown: Number of items per exam with p beyond recommended values of 0.4 and 0.8 respectively according to algorithm applied

Total item correlation r

Analysis of total item correlations showed no clear advantage for a particular algorithm.

Mean total item correlations at most differed by 0.03 between the three scoring algorithms with a slight tendency toward higher values when applying PS_{50} and $PS_{1/m}$, yet fewer items yielded r below 0.20 or 0 when applying partial scoring (see Table 4). A maximum of one item had negative r values per test examined, once for DS, once for PS_{50}.

Performance

In all six tests examined the average score when dichotomously scored was 4–5 points below that of scenarios with partial credits (Δ PS_{50}−DS: min = 4.30, max = 6.19, M = 4.92 points, SD = 0.69), with $PS_{1/m}$ regularly allowing for one extra point in comparison to PS_{50} (Δ $PS_{1/m}$−PS_{50}: min = 0.81, max = 1.70, M = 1.05 points, SD = 0.33), details shown in Table 5.

Table 4 Unadjusted total item correlations r and index of discrimination D for Pick-N items according to algorithm used

Term	n (items)	M(r), (SD)			r < 0.2		
		DS	PS_{50}	$PS_{1/m}$	DS	PS_{50}	$PS_{1/m}$
Summer 05	30	0.27 (0.10)	0.29 (0.11)	0.29 (0.11)	9	7	7
Winter 05/06	30	0.27 (0.13)	0.27 (0.11)	0.28 (0.10)	7	6	7
Summer 06	31	0.30 (0.10)	0.31 (0.11)	0.31 (0.11)	5	5	6
Winter 06/07	28	0.35 (0.10)	0.36 (0.10)	0.36 (0.10)	3	2	2
Summer 07	27	0.32 (0.10)	0.34 (0.12)	0.33 (0.11)	5	4	6
Winter 07/08	22	0.32 (0.12)	0.35 (0.13)	0.35 (0.12)	3	3	2
Mean	28	0.30 (0.11)	0.32 (0.11)	0.32 (0.11)	5.3	4.5	0.5

Standard deviation in parenthesis. Also shown: Number of items with r below 0.2 according to algorithm applied

Table 5 Test performance; average points awarded on Pick-N subsets depending on algorithm used; standard deviation (*SD*) in brackets

Term	n (items)	DS Points earned (SD)	PS_{50}	$PS_{1/m}$
Summer 05	30	19.4 (3.4)	24.2 (2.1)	25.1 (1.9)
Winter 05/06	30	15.7 (3.9)	21.9 (2.2)	23.6 (2.0)
Summer 06	31	20.0 (4.0)	25.1 (2.3)	26.0 (2.0)
Winter 06/07	28	17.3 (4.3)	21.9 (2.7)	22.8 (2.3)
Summer 07	27	17.5 (3.8)	21.8 (2.3)	22.6 (1.9)
Winter 07/08	22	12.1 (3.1)	16.5 (2.0)	17.6 (1.7)
Mean	28	17.0 (3.7)	21.9 (2.3)	23.0 (2.0)

Conclusions

Ripkey's findings regarding the differences between dichotomous and partial crediting in Pick-N multiple best select items have been reproduced in this study (Ripkey et al. 1996) in a summative, undergraduate setting. Differences between the two partial credit algorithms PS_{50} and $PS_{1/m}$ were marginal and resulted in similar psychometric data. Spearman-Brown prediction of test reliabilities showed little difference in the different scoring settings when projected to 100 item exams.

Comparing item and test attributes with those of A-type single best answer exams further suggest Pick-N items to be an apt addition to the assessment toolbox.

Discussion

A threshold as used in the algorithm PS_{50} (i.e. 50% of true answers in a Pick-N item have to be correctly identified to get credit for that item at all) could separate guesses based on partial knowledge (tending towards blind guesses), from those based on more complete knowledge (educated guesses as needed in clinical routine). Exactly where that threshold should be for Pick-N MCQs in norm-oriented scoring should be subject of future studies in this field.

To what extent cueing effects (unintended hints) in Pick-N tests differ from A-type MC exams so far remains unknown. Haladyna finds three answer options for single best answer items to be good enough when well reflected upon as examinees often only ever really

ponder few distractors offered (Haladyna and Downing 1993), a notion later supported in Rodriguez's meta-analysis in 2005 (Rodriguez 2005). How many answer possibilities and distractors make a good Pick-N item has so far not been studied. It is, however, assumed that answer choices and plausible distractors have to be just as well chosen as in any other MC format. Given that test performance tends to increase just by awarding partial credit instead of scoring items dichotomously the adjustment of pass and grade thresholds should be subject to discussion in individual faculties.

Taking all this into consideration the Pick-N format seems a multiple choice format for summative undergraduate assessment worthwhile considering in medical education when allowing for partial crediting. In the end, examiners have to decide for themselves, whether or not to make use of Pick-N MCQs and how to score them, as long as the rationale behind these rulings remains transparent. One should bear in mind the implications of the algorithm used on psychometrics, scores and student satisfaction. Further validation of Pick-N items and partial scoring algorithms for the Pick-N format in relation to other formats and contexts of assessment is encouraged.

Acknowledgments The authors would like to extend their gratitude towards René Krebs of Berne University, Switzerland and Andreas Möltner of University of Heidelberg, Germany for their helpful suggestions.

Appendix

Exemplary Pick-N Item, taken from the mid-term exam of 2007/2008:

Mrs W., a 39 year old patient who up until then had been in good health, presents herself as an in-patient complaining of a persistent cough. Three weeks ago she had suffered from a cold and initially, the cough was non-productive. Now, in addition to the cough that has been producing yellow sputum for 3 days, her temperature has risen to 39°C and she complains of stabbing pain on the left when inhaling.

Which of the following initial diagnostic steps do you take on the basis of a suspected case of ambulatory pneumonia in addition to the physical examination?

(*Select **four** answers*!)

A. Blood gas analysis
B. Bronchoscopy
C. Thorax CT
D. Thorax x-ray
E. Standard lab
F. Spiroergometry
G. Sputum tests
H. Thorax sonography
I. Ventilation scintigraphy

Answers A, D, E and H had been identified as correct answers

This item would be scored as follows:

m	0	1	2	3	4
DS	0.00	0.00	0.00	0.00	1.00
P_{50}	0.00	0.00	0.50	0.50	1.00
$P_{1/m}$	0.00	0.25	0.50	0.75	1.00

Points earned in Pick-N item with n = 4 correct answers. m denominates the number of correct answers chosen by candidate

References

Albanese, M. A. (1993). Type K and other complex multiple-choice items: An analysis of research and item properties. *Educational Measurement: Issues and Practice, 12*(1), 28–33.

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

Ben-Simon, A., Budescu, D. V., & Nevo, B. (1997). A comparative study of measures of partial knowledge in multiple-choice tests. *Applied Psychological Measurement, 21*(1), 65–88.

Beullens, J., Struyf, E., & van Damme, B. (2005). Do extended matching multiple-choice questions measure clinical reasoning? *Medical Education, 39*(4), 410–417.

Beullens, J., Struyf, E., & van Damme, B. (2006). Diagnostic ability in relation to clinical seminars and extended-matching questions examinations. *Medical Education, 40*(12), 1173–1179.

Bland, A. C., Kreiter, C. D., & Gordon, J. A. (2005). The psychometric properties of five scoring methods applied to the script concordance test. *Academic Medicine, 80*(4), 395–399.

Case, S. M., & Swanson, D. B. (1993). Extended-matching items: A practical alternative to free-response questions. *Teaching and Learning in Medicine, 5*(2), 107–115.

Case, S. M., & Swanson, D. B. (2001). *Constructing written test questions for the basic and clinical sciences*. Philadelphia: National Board of Medical Examiners.

Coderre, SP., Harasym, P., Mandin, H., & Fick, G. (2004). The impact of two multiple-choice question formats on the problem-solving strategies used by novices and experts. *BMC Medical Education*. 4(23).

Downing, S. M. (2004). Reliability: On the reproducibility of assessment data. *Medical Education, 38*(9), 1006–1012.

Epstein, R. M. (2007). Assessment in medical education. *New England Journal of Medicine, 356*, 387–396.

Fournier, J. P., Demeester, A., & Charlin, B. (2008). Script concordance tests: Guidelines for construction. *BMC Medical Informatics and Decision Making*. 8(18).

Haladyna, T. M., & Downing, S. M. (1989a). A taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education, 2*(1), 37–50.

Haladyna, T. M., & Downing, S. M. (1989b). Validity of a taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education, 2*(1), 51–78.

Haladyna, T. M., & Downing, S. M. (1993). How many options is enough for a multiple-choice test item? *Educational and Psychological Measurement, 53*(4), 999–1010.

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

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

Kassirer, J. P., & Kopelman, R. I. (1991). *Learning clinical reasoning*. Baltimore: Williams & Wilkins.

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

Lord, F. M. (1963). Formula scoring and validity. *Educational and Psychological Measurement, 23*(4), 663–672.

Möltner, A., Schellberg, D., & Jünger, J. (2006). Grundlegende quantitative Analysen medizinischer Prüfungen. *GMS Zeitschrift für Medizinische Ausbildung*. 23(3).

Nendaz, M. R., & Tekian, A. (1999). Assessment in problem-based learning medical schools: A literature review. *Teaching and Learning in Medicine, 11*(4), 232–243.

Norcini, J. J., Swanson, D. B., Grosso, L. J., Shea, J. A., & Webster, G. D. (1984). A comparison of knowledge, synthesis, and clinical judgment: Multiple-choice questions in the assessment of physician competence. *Evaluation & the Health Professions., 7*(4), 485–499.

Ripkey, D. R., Case, S. M., & Swanson, D. B. (1996). A "new" item format for assessing aspects of clinical competence. *Academic Medicine, 71*(10), S34–S36.

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

Rotthoff, T., Baehring, T., Dicken, H. D., Fahron, U., Richter, B., Fischer, M., & Scherbaum, W. (2006). Comparison between long-menu and open-ended questions in computerized medical assessments. A randomized controlled trial. *BMC Medical Education* 6(50).

Schuwirth, L. W. T., & van der Vleuten, C. P. M. (2003). ABC of learning and teaching in medicine: Written assessment. *British Medical Journal, 326*, 643–645.

Swanson, D. B., Holtzman, K. Z., & Allbee, K. (2008). Measurement characteristics of content-parallel single-best-answer and extended-matching questions in relation to number and source of options. *Academic Medicine, 83*(10), 21–24.

Wakeford, R. E., & Roberts, S. (1984). Short answer questions in an undergraduate qualifying examination: A study of examiner variability. *Medical Education, 18*(3), 168–173.