Note relative au système de « points négatifs » associé aux évaluations de type « questionnaires à choix multiple » au sein de l'UCL.

Fait à Bruxelles le 19 juin 2016,

Chères conseillères, chers conseillers,

Depuis de très nombreuses années, du fait du nombre important d'étudiant-e-s, de nombreux, enseignants, au sein des différentes facultés de l'UCL ont fait le choix d'évaluer les acquis des étudiant-e-s via un système d'évaluation de type « questionnaire à choix multiple ».
De peur que ce système d'évaluation ne soit biaisé par une réponse aléatoire au questionnaire, certains enseignants ont pris la décision d'introduire un système de pénalité sous la forme de « points négatifs »[1] (-0.25, -0,5, -1, etc…) en cas d'une mauvaise réponse ou d'une abstention.
Ce système de pénalités a été et est actuellement remis en cause dans diverses institutions.

En 2014, à l'issue des résultats d'une « review » publiée par des chercheurs de l'Université de Gand[2] ayant analysé la littérature concernant cette problématique et de « recherches quasi-expérimentales » menées au sein de cette institution[3] , cette dernière, a pris la décision de remplacer les **« points négatifs »** en **échange** d'un système de **« standard setting »**.

Tout d'abord, qu'est-ce que le « standard setting » ?

Le « standard setting » un système, qui est déjà utilisé par un certains professeurs de l'UCL. Il consiste instaurer un « cut-off point » supérieur aux traditionnels 50% afin de compenser les réponses ayant potentiellement été données aléatoirement par les étudiants.
Ce « cut-off point » est déterminé par l'enseignant à un seuil pouvant souvent correspondre à 55 ou 60% de réponses justes.
Donc en somme, il faut répondre juste à plus de 50% des questions pour réussir l'évaluation.

Quels ont été les résultats retenus qui ont poussé l'Université de Gand à tourner le dos au système de points négatifs ?

1. L'Université de Gand a appliqué différentes méthodes d'évaluation de manière répétée demandant aux étudiant-e-s de déterminer laquelle était la plus optimale.
2. De nombreux enseignants appliquaient un système de « points négatifs » tout simplement car il y étaient habitués.
3. Avant, durant et après l'évaluation les étudiants passent beaucoup de temps sur des considérations stratégiques.
4. Des étudiant-e-s avec une même maîtrise de la matière diffèrent de manière importante lorsqu'ils doivent choisir de répondre ou de ne pas répondre.
5. La décision de répondre ou de ne pas répondre n'est pas toujours prise de manière rationnelle.

---

[1] Points négatifs relatifs à une question. (ne concerne pas les penalités attribuées en cas d'échecs à des parties de l'enseignement qui est évalué)
[2] E. Lesage et al. / Studies in Educational Evaluation 39 (2013) 188–193
[3] http://www.ugent.be/en/education/degree/practical/studentadmin/OEREnglish/multiplechoice.htm

6. Les différences relatives au « guessing behavior » (la tendance à essayer de deviner la réponse) ont une influence significative au niveau de la note finale des étudiant-e-s.
7. Les étudiant-e-s bénéficiant le plus de la transition d'un système de « points négatifs » vers un système de « standard setting » sont ceux qui ont le moins tendance à essayer de deviner la réponse.
8. Les chances de réussir en devinant sont aussi faibles avec le système de « points négatifs » qu'avec le système de « standard setting ».

En somme l'Université de Gand a abandonné le système de « points négatifs » en raison des considérations « stratégiques » que ce système induisait et des variations relatives fait que les étudiant-e-s n'ont pas tou-te-s la même personnalité ce qui induit que les étudiant-e-s ne réagissent pas tous de la même manière face à une même évaluation. Cette transition de système, n'induit pas de différence dans la difficulté de l'évaluation. Le but est que l'étudiant-e se concentre au maximum sur l'étude et l'apprentissage des acquis d'apprentissage et ne se laisse plus distraire par des considérations stratégiques.

Quid de l'UCL ?

Que trop bien conscients de l'imperfection du système d'évaluation de type « QCM » mais aussi du fait qu'il soit très difficile de s'en passer dans les études où les cohortes d'étudiant-e-s sont importantes. C'est pourquoi il semble d'autant plus important de l'améliorer.
C'est dans cette vision qu'il semble plus qu'opportun de nous inspirer de l'Université de Gand en demandant à l'UCL (via les différentes instances où l'AGL a un pouvoir de parole) :

De demander aux enseignants ayant recours à un système de « points négatifs » dans leurs « QCM » à passer à un système de « standard setting ».
D'abolir ainsi la possibilité pour un enseignant d'user de «points négatifs » dans ses « QCM ».

La suppression de ce système de « points négatifs » au bénéfice d'un système de « standard setting » permettrait d'éliminer les biais (personnalité et de stratégie cités plus haut dans le texte) et d'ainsi **améliorer l'évaluation, la sélection**[4] et **l'équité des étudiant-e-s face à l'évaluation.**
Cette transition améliorerait donc directement la qualité des diplômes délivrés par l'UCL.

«*Pour s'améliorer il faut changer, pour être parfait il faut avoir souvent changé. »*
W.Churchill.

Salim abene
- Conseiller AGL
- Délégué 2015-2016 du BDE-MEDECINE en commission d'enseignement de la faculté de médecine
-Délégué de cours de 3ba Médecine 2015-2016
Mail : Salim.abene@student.uclouvain.be

[4] sélection restreignant l'accès à différentes filières, cycle/bloc supérieur, stages, formations à l'étranger, attestations…

Studies in Educational Evaluation 39 (2013) 188–193

Contents lists available at ScienceDirect

Studies in Educational Evaluation

journal homepage: www.elsevier.com/stueduc

Scoring methods for multiple choice assessment in higher education – Is it still a matter of number right scoring or negative marking?

Ellen Lesage *, Martin Valcke [1], Elien Sabbe [2]

Zomerstraat 60, 9000 Ghent, Belgium

ARTICLE INFO

Article history:
Received 31 October 2012
Received in revised form 7 July 2013
Accepted 8 July 2013

Keywords:
Assessment
Higher education
Multiple choice

ABSTRACT

In higher education, a multiple choice test is a widely known format for measuring student's knowledge. The debate about the two most commonly used scoring methods for multiple choice assessment – number right scoring (NR) and negative marking (NM) – seems to be a never-ending story. Both NR scoring as NM do not seem to meet the expectations. However, available research hardly offers alternative methods. Clearly, there is a growing need to explore these alternative scoring methods in order to inform and support test designers. This review aims to present an overview of (alternative) scoring methods for multiple choice tests, in which strengths and weaknesses of each method are provided.

Contents

Purpose of this review . . . 188
Methodology . . . 189
Conventional scoring methods . . . 189
Number right scoring . . . 189
Negative marking . . . 189
A preliminary conclusion about number right versus negative marking . . . 190
Non-conventional scoring methods . . . 190
Scoring methods that acknowledge partial mastery . . . 190
Multiple answers . . . 190
Retrospective correcting for guessing . . . 191
Standard setting . . . 191
Conclusions . . . 191
References . . . 192

Purpose of this review

The major goal of testing is to extract student's knowledge of the subject matter from responses to tests. The test score has to correspond as close as possible to the 'true' level of knowledge mastery of students. The debate about scoring multiple choice test formats presents a particular case in the research literature. Many studies have compared scoring methods, however, the comparison involved tests from varying knowledge domains, neglected differences in test length and/or applied different analysis criteria. As a consequence, comparisons of multiple choice test approaches turned out to be difficult and presented mixed results. In addition, the research conclusions neglected the particular impact of internal and external variables. The specific educational context is a key external variable. Internal variables are linked to student variables such as risk-taking behavior and test anxiety. Not surprisingly, an unequivocal, evidence-based advice concerning the scoring of multiple-choice tests does yet not exist.

In the present paper, we do not continue the debate whether the two most commonly scoring methods, number right (NR) or negative marking (NM), is to be preferred. We rather aim at

* Corresponding author. Tel.: +32 477455752.
E-mail addresses: ellenlesage@hotmail.com (E. Lesage), Martin.Valcke@UGent.be (M. Valcke), Elien.Sabbe@UGent.be (E. Sabbe).
[1] Tel.: +32 92648675.
[2] Tel.: +32 93310062.

0191-491X/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.stueduc.2013.07.001

presenting a comprehensive review of multiple choice scoring methods currently used in higher education, resulting in an analysis of weaknesses and strengths. Appendix 1 summarizes the multiple choice scoring methods, commonly used in previous studies from 1979 onwards. We compare these scoring methods along several dimensions that are important to multiple choice tests: scoring rule, ideal answer strategy, reliability and validity, students' guessing behavior, partial knowledge and educational issues.

Methodology

The online databases of ISI Web of Science, the Educational Resources Information Center (ERIC), Science Direct and Google Scholar were searched between the years 1979 and 2011. The following keywords were used: *multiple choice*; *assessment*; *examination*; *higher education*; *scoring method*; *number right scoring*; *negative marking*; *formula scoring*; *correction for guessing*; *standard setting*; *guessing or risk-taking behavior*. All abstracts were reviewed and relevant documents were selected. These relevant articles were reviewed and their references were searched for associated articles.

In the context of the review of the research, in particular the knowledge domains of the medical sciences was covered, and this in relation to the topics standard setting and scoring methods that acknowledge partial mastery. In addition, specially the knowledge domains psychology, economics, law and mathematics appeared in the review articles.

Conventional scoring methods

Number right scoring

Traditionally, multiple choice tests have been scored using a conventional number right (NR) scoring method (Bereby-Meyer, Meyer, & Flascher, 2002; Kurz, 1999). Correct answers are scored with a positive value, incorrect answers and absent or omitted answers with a value of zero. The sum of the scores for correct responses is the test score. A major concern about this scoring method, is that students can answer correctly through guessing (Choppin, 1988). Students without the ability to solve a particular item, gain marks by guessing (Budescu & Bar-Hillel, 1993; Frary, 1988; Kubinger, Holocher-Ertl, Reif, Hohensinn, & Frebort, 2010). Guessing introduces a random factor into test scores that lowers reliability and validity (Bereby-Meyer et al., 2002; Burton, 2001; Kubinger et al., 2010; Prihoda, Pinckard, McMahan, & Jones, 2006). Test designers can as such not distinguish between correct answers based on knowledge mastery versus those based on a guess (Bar-Hillel, Budescu, & Attali, 2005).

Negative marking

Various scoring formulas have been presented to correct for guessing (Kurz, 1999). The 'rights minus wrongs' correcting model (Kurz, 1999) is predominant and penalizes the student for incorrect responses. The fundamental idea behind this scoring method, is that students acknowledge they will lose marks for incorrect answers (Betts, Elder, Hartley, & Trueman, 2009). As a consequence, students are discouraged to guess, and this is expected to increase test reliability and validity because the test score is a truer reflection of a student's ability (Kurz, 1999). Comparable to NR, a correct response results in a positive score and omitted items result in no mark (Hammond, McIndoe, Sansome, & Spargo, 1998). A straightforward interpretation is that the expected total score should be zero if a student guesses all answers at random. For this to happen, the penalty for an incorrect answer should be $1/(n-1)$, where n stands for the number of choices (Karandikar, 2010). We use the term 'negative marking' to describe this scoring method.

An alternative other model, proposed by Traub, Hambleton, and Singh (1969), rewards a student for not guessing by awarding points for omitted items rather than penalizing for incorrect responses. This presents a psychological advantage since it rewards the desired behavior rather than penalizing undesirable behavior (Crocker & Algina, 1986). Prieto and Delgado (1999) favor the latter scoring rule as the best of the scoring methods discussed thus far, and refer to resulting performance indicators and score reliability. Students do not feel threatened by receiving a reward for skipping items, as compared to receiving a penalty for incorrect responses.

Studies (Burton, 2002; Muijtjens, van Mameren, Hoogenboom, Evers, & van der Vleuten, 1999) report an increase in validity or reliability when negative marking is implemented. However, these studies only show slight improvements (Kurz, 1999) and they specifically examine true/false/items. Because of the high susceptibility to guessing, the results cannot simply apply to items with more alternative choices. Moreover, there is no consistency concerning applied analysis criteria, test length, knowledge domains and test instructions in these studies.

There remain other concerns about the use of negative marking. The rationale behind this scoring method – discouraging guessing behavior – seems to miss its purpose. It does not solve the guessing problem (Bar-Hillel et al., 2005; Betts et al., 2009); it even introduces new problems that are not observed in the case of NR scoring. Students differ in their guessing behavior: some students dare to take more risks than others. This introduces a first concern about students' risk attitudes adding to uncontrolled sources of variance, thus reducing the test's reliability and validity (Bar-Hillel et al., 2005). Moreover, guessing might benefit students who guess frequently compared to students with equal ability levels who do not guess (Choppin, 1988; Muijtjens et al., 1999). Some authors argue that by implementing negative marking, multiple choice tests rather measure students' answering strategies and risk-taking behavior instead of the mastery of domain knowledge (Budescu & Bar-Hillel, 1993; Choppin, 1988; Fowell & Jolly, 2000; Hammond et al., 1998; Kurz, 1999; Moss, 2001; Prihoda et al., 2006).

A third concern is related to the instruction to be given to students about guessing behavior. When negative marking was first introduced, students were simply advised not to guess (Davis, 1967; Frary, 1988). Change came when students were instructed to guess, whenever they could eliminate at least one or more alternative choices (Betts et al., 2009; Davis, 1967; Frary, 1988; Hammond et al., 1998). The study of Betts et al. (2009) suggests that guessing does not significantly reduce student performance. Hammond et al. (1998) advise instructors to be very cautious about instructing students not to guess. However, Budescu and Bar-Hillel (1993) point out the contradictory character of these instructions to guess, since the underlying principle of this formula is to discourage guessing. These authors conclude it is almost impossible to give recommendations that are fair and beneficial to all students. Students react in inconsistent ways, thus the question whether we have to instruct a student to guess or not is therefore far more difficult to answer than it seems (Budescu & Bar-Hillel, 1993). In this respect, it is also difficult for students to figure out the optimal decision strategy under negative marking.

Finally, there is confusion about the amount of the penalty given for incorrect answers (Karandikar, 2010). Some authors (Budescu & Bar-Hillel, 1993; Espinosa & Gardeazabal, 2010) state that an effective penalty that effectively discourages guessing, should exceed the standard penalty of $1/(n-1)$.

A preliminary conclusion about number right versus negative marking

Both scoring methods affect in a particular way the reliability of the test. Guessing might affect test reliability (Choppin, 1988; Ng & Chan, 2009). And though negative marking is expected to discourage guessing, it wrestles in the end with a comparable impact on test reliability due to students' guessing behavior. This results in the position of authors to abandon scoring methods that penalize for incorrect responses and to apply the simple number right scoring (Abu-Sayf, 1979; Kurz, 1999).

A possible solution to avoid negative marking and minimize the effect of guessing on test reliability, could be to increase the number of items in the test, as well as the number of alternative choices in multiple choice questions, to increase test reliability (Burton, 2001, 2004; Karandikar, 2010; Muijtjens et al., 1999; Zimmerman & Williams, 2003). Although the average time to complete a test is higher under negative marking than under number right scoring (Ben-Simon, Budescu, & Nevo, 1997; Kurz, 1999), increasing the number of both items and alternative choices does not always seem feasible, especially for university or high school examinations where tests are timed (Burton, 2001; Muijtjens et al., 1999). Developing more item choices does not always result in a decrease in guessing as test designers meet additional difficulties when they have to develop extra item choices. It is very likely that these extra alternatives hardly function as effective distractors and as a consequence they miss their purpose. In contrast when proper alternatives can be found, the chance of guessing the correct answer, decreases.

From the perspective of test designers, a choice is to be made between striving for high reliability on the base of negative marking or avoiding bias due to risk-taking behavior in a number right situation (Muijtjens et al., 1999). As discussed above, making this choice might be 'a mission impossible'. We therefore move to another approach and look for alternative scoring methods.

Non-conventional scoring methods

Scoring methods that acknowledge partial mastery

A concern with both number right scoring and negative marking, is that they do not take into account partial knowledge mastery (Bradbard, Parker, & Stone, 2004; Kurz, 1999). Although students cannot always identify the correct answer, they can often determine that some of the choice options are clearly incorrect. This is labeled as 'partial knowledge' (Coombs, Milholland, & Wome, 1955; Frary, 1980). A way to overcome this shortcoming, is to develop methods that allow a more accurate measurement of student knowledge (Ben-Simon et al., 1997; Frary, 1989). This results in what Frary (1989) calls 'partial-credit scoring methods'. These are scoring methods that attempt to capture information about a student's degree of level of knowledge with respect to each choice presented in relation to a test item. NR scoring only discriminates between full knowledge and absence of knowledge, whereas NM can only distinguish between full knowledge, misinformation and absence of knowledge (Ben-Simon et al., 1997). Although a variety of different partial-credit scoring methods exists, we can distinguish three main formats:

(1) The *liberal multiple-choice test* allows students to select more than one answer to a question if they feel uncertain which alternative is correct (Bush, 2001).
(2) Students are instructed to cross out all alternatives they consider to be incorrect: *elimination testing* (ET) (Kurz, 1999).
(3) In *confidence weighting* (CW), students have to indicate what they believe is the correct answer and how confident they are about their choice (Kurz, 1999).

The score obtained on a liberal multiple choice test is likely to be a better reflection of a student's true knowledge mastery as compared to the score obtained on the base of a conventional multiple choice test version based on the same questions (Bush, 1999, 2001). Students are able to express explicitly their partial knowledge mastery (Bush, 2001). Compared to NR scoring, students are likely to obtain a higher score whenever they can eliminate two or more incorrect options (Jennings & Bush, 2006). Ben-Simon et al. (1997) compared seven different scoring methods awarding partial credits. None of the methods emerged as the best, considering test validity and reliability. However, ET tended to discriminate to a better extent between students than NR and NM and reflected a more optimal reliability as compared to NR and NM (Ben-Simon et al., 1997; Bradbard et al., 2004). ET tests help to distinguish between full and partial knowledge mastery (Bradbard et al., 2004). Some authors studied students' attitudes toward partial-credit scoring methods (Ben-Simon et al., 1997; Bush, 1999; Gardner-Medwin, 1995). The results of these studies show that – initially – the principles of a partial-credit scoring system are complicated for students. But, with experience, they no longer report problems (Ben-Simon et al., 1997; Bush, 1999; Gardner-Medwin, 1995). Moreover, students appreciate these non-conventional marking schemes, once they are familiar (Bush, 1999).

Partial-credit scoring methods do not seem to affect test reliability as compared to NR scoring (Alnabhan, 2002; Bradbard et al., 2004). In addition, these scoring methods have to be approached with care. Bush (2001) notes that no partial-credit scoring method has been identified as superior compared to the conventional multiple choice scoring methods. Also, the complexity of both test-taking and test-scoring can be a possible reason why these alternative methods are not widely used (Bradbard et al., 2004; Kurz, 1999). Research whether these alternative methods discourage guessing, is also scarce. To what extent is it certain students will follow the instructions related to, e.g., confidence testing? Is it plausible to assume that risk-taking students will also take the risk to indicate that they feel more confident about their response than they actually do? Also, partial-credit scoring methods seem to be more appropriate in some situations than others (Bush, 2001). ET scoring and liberal multiple choice tests might be useful in content areas where partial or full misinformation results in life-threatening consequences, such as medical training for doctors, nurses or pharmacists (Bradbard et al., 2004; Frary, 1989; Jennings & Bush, 2006).

Despite the fact that partial credit scoring methods can measure partial knowledge, we still question whether such methods really discourage guessing. Even though they award students for partial knowledge, they offer no solution for guessing. Risk-taking students will again not behave as ideal test takers and will bend the rules of the partial-credit scoring method (Rogers, 1999). In a liberal multiple choice test, a student can, e.g., select only one choice, even though he thinks another choice might also be correct. Students can again easily start guessing and thus increase their test scores. We again have to conclude that this scoring method might also miss the purpose of countering guessing behavior.

Multiple answers

In most cases, multiple choice tests consist of single-answer items (SA) but also a variety of multiple answer items (MA) does occur. In some MA formats students are instructed that each item presents at least one correct choice (Kurz, 1999), in others students are told how many answers are correct (Kubinger et al., 2010). Items can be scored as solved only if all correct response options are marked, but none of the incorrect others, and in other cases incorrect options that are marked can lead to negative scores.

In a recent study of Kubinger et al. (2010), a multiple choice format of two solutions and three distractors (2 of 5) was used in which students knew that two options are correct and three are incorrect. An item is scored correctly if both correct answers are marked and none of the other incorrect options are marked. This format was identified as more difficult than the format with one solution and five distractors (1 of 6). The authors consequently assumed that the guessing effect significantly decreased by using the '2 of 5'-format.

Thayn (2011) compared the discriminating power, item difficulty and item reliability of SA and MA items. Students were told how many options they had to select. Each item is scored dichotomously, with one point awarded if the correct answer is selected and zero points if any of the incorrect answers are selected. Both MA and SA items showed almost identical statistical performance characteristics. These results provide strong evidence that MA items perform at least as well as SA items (Thayn, 2011).

Nevertheless, also the MA format has been criticized because test items (a) tend to become more difficult to solve, (b) reflect a lower discriminating power and (c) take more time to answer (Thayn, 2011). Due to the increase in time to answer, test designers can present fewer test items. This implies that with less questions being presented, one can nevertheless cover the same content span as a test consisting of far more single-answer items. MA items – assuming especially those items where students do not know the number of correct answers – also have drawbacks related to the possible ambiguity in student interpretation of items. It is likely to believe that students read MA items more carefully than SA items. But the question can be posed whether it is feasible to develop every multiple answer item in an unequivocal way? Since MA items that penalize incorrect answers or do not inform students about the number of correct answers generate similar levels of uncertainty as negative marked items do, it is reasonable to assume that they are influenced by students' guessing behavior as well. Therefore, the development of MA items should be well considered.

Retrospective correcting for guessing

Another scoring method for multiple choice tests, is the retrospective correcting for guessing format. According to this format, an absent answer is seen as an incorrect answer. So for their own benefit, students need to give an answer to every question – even if they do not know the correct answer – since the expected score for responding is likely to be higher than omitting. The correction for guessing is implemented afterwards – or retrospectively – so we use the term retrospective correcting for guessing. These corrected scores are based on an estimation of the student's guessing behavior.

Prihoda et al. (2006) compared scores under number right scoring and retrospectively corrected scores on multiple choice tests with scores on short-answer tests. Compared to the number right scores, corrected scores agreed more with the short-answer scores and so indicating an increased validity. Scharf and Baldwin (2007) compared three different scoring methods: one with zero, an intermediate and a maximum penalty. They concluded that the intermediate option or the retrospective correcting for guessing is the most justifiable. It penalizes blind guessing – which is an advantage compared to the scoring method with zero penalty – and partial knowledge lessens the negative impact on the final score in the end – which is an advantage compared to the scoring method with a maximum penalty. However, their conclusions are based on pure mathematical analyses on raw data and not on empirical findings.

In contrast to negative marking, risk attitudes are excluded as every student has to answer all questions. Students are forced to guess, which can be difficult to justify (Kurz, 1999). So the question rises whether this format is a good alternative for number right scoring and negative marking. In content areas where students have to know the answer, e.g., medical training for doctors, this scoring method is definitely not appropriate.

Standard setting

Test designers can also vary the standards used when scoring multiple choice tests. Independent of the issue of guessing in multiple choice exams, two main options can be distinguished: norm-referenced versus criterion-referenced assessment. The *norm-referenced assessment* approach builds on a relative standard that reflects the particular performance of the current group of students participating in the test (Bandaranayake, 2008). As a result, these standards vary depending on group differences (van der Vleuten, 2010). Most test designers tend not to adopt a norm-referenced approach (de Gruijter, 2001). *Criterion-referenced assessment* is based on an absolute standard that assesses students against a specified achievement level (Dunn et al., 2007). This assessment method ensures the fact that students achieve a specified competence level (Bandaranayake, 2008), which is known in advance (de Gruijter, 2001). In defining the standard, the effect of guessing can be taken into account. For example, the standard of a multiple choice exam consisting of 40 items with 4 choices can be set at 50%. Since 10 of those items however will generally be answered correctly when students randomly guess, the passing score can be raised accordingly to 25 out of 40 items. Since an absent answer is seen as an incorrect answer, students are better off giving an answer to all of the questions, rather than omitting. Unfortunately, a criterion-referenced method is affected by variations in test difficulty (Muijtjens et al., 1999; van der Vleuten, 2010).

As both approaches struggle with criticisms, Cohen-Schotanus and van der Vleuten (2010) propose an alternative approach, which combines a pre-fixed cut-off score with a norm-referenced standard, focusing on the high performers as a relative point of reference (Cohen-Schotanus & van der Vleuten, 2010). Dochy, Kyndt, Baeten, Pottier, and Veestraeten (2009) compared different standard setting methods, including Cohen's method. Although this method scored reasonably well, the standard might still depend on the knowledge level of the particular group of students. The most appropriate standard setting method for all tests does not exist (Dochy et al., 2009).

A general concern of standard setting is the process of determining the marks, which is rather complicated. Also, one can question the process of setting a passing score or cut-off point (Bandaranayake, 2008; Downing, 2004; Downing, Lieska, & Raible, 2003; Norcini, 2003). How does one know when a certain standard is acceptable? Finally, test designers can have some difficulties to accept the fact that students are forced to guess. However – comparable to retrospective correcting for guessing – students' guessing behavior does not play a role.

Conclusions

Multiple-choice tests can be scored in a variety of ways. For a long time, the debate has been single-sided and mainly focused on number right scoring versus negative marking. Above, advantages and disadvantages of both methods have been compared. Since no empirical evidence is available in the literature that helps to direct the choice between either approaches, we argued to focus on alternative approaches.

Since the literature about these alternatives does not present a clear direction in view of opting for the most optimal solution, we

can only conclude there is a growing need to evaluate alternatives in valid and educational settings. Thus far, test designers hardly dare to experiment with these alternatives. Therefore research is needed to reduce the gap between theoretical possibilities and practice. It is important to inform test designers about these scoring methods and their respective advantages or disadvantages. It can help them to take legitimate decisions, ideally in dialog with key actors involved in an educational program. In this respect, more consistency in scoring methods of multiple choice tests is recommended at program or institutional level in higher education, as students are also expected to benefit from a coherent scoring approach.

Appendix 1. Overview of scoring methods for multiple choice tests

	Number right scoring	Negative marking	Reward for omissions	Partial credit scoring
Scoring rule	Incorrect answers and omissions are not penalized	Incorrect answers are penalized. Omissions are not penalized	Omissions are rewarded	Rewards for partial knowledge. Different partial-credit scoring methods exist
Ideal answer strategy	Always guess	If you can eliminate at least one answer choice, you should guess	If you can eliminate at least one answer choice, you should guess	If you can eliminate at least one answer choice, you should guess
Reliability and validity	Decreased reliability due to guessing	Small increase of reliability and validity, however results are inconsistent	Inconsistent results	Inconsistent results
Students' guessing behavior	If students answer all items, there is no effect of guessing behavior	Students' guessing behavior adds a source of variance	Students' guessing behavior adds a source of variance	Students' guessing behavior adds a source of variance
Partial knowledge	Not taking into account	Not taking into account	Not taking into account	Takes partial knowledge into account, if a student does not guess
Educational issues	Encourages students to guess and rewards it	Confusion about the instruction for guessing. Does not discourage guessing	Rewarding omissions seems more appropriate than penalizing incorrect answers. Does not discourage guessing	Test-taking and test scoring is more complex. Does not discourage guessing

	Multiple-answer items	Retrospective correcting for guessing	Standard setting
Scoring rule	Every choice option becomes a true/false item. Varies from dichotomous to partial-credit scoring	A retrospective correcting for guessing is subtracted from the sum of the correct answers. Omissions are scored as incorrect	A standard for minimal passing is set. Omissions are scored as incorrect
Ideal answer strategy	None. Students should know the number of correct answers	Always guess	Always guess
Reliability and validity	Inconsistent results	Research is scarce	Research is scarce
Students' guessing behavior	Students' guessing behavior hardly adds a source of variance	If students answer all items, there is no effect of guessing behavior	If students answer all items, there is no effect of guessing behavior
Partial knowledge	With a partial-credit scoring, partial knowledge would be taking into account	Not taking into account	Not taking into account
Educational issues	Developing univocal MA items requires an enormous effort for test designers	Forced guessing is difficult to justify	Forced guessing is difficult to justify. One can question the process of determining the standard

References

Abu-Sayf, F. K. (1979). The scoring of multiple choice tests: A closer look. *Educational Technology, 19*, 5–15.

Alnabhan, M. (2002). An empirical investigation of the effects of three methods of handling guessing and risk taking on the psychometric indices of a test. *Social Behavior and Personality, 30*(7), 645–652.

Bandaranayake, R. C. (2008). Setting and maintaining standards in multiple choice examinations: AMEE Guide No. 37. *Medical Teacher, 30*, 836–845.

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

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.

Bereby-Meyer, Y., Meyer, Y., & Flascher, O. M. (2002). Prospect theory analysis of guessing in multiple choice tests. *Journal of Behavioral Decision Making, 15*, 313–327.

Betts, L. R., Elder, T. J., Hartley, J., & Trueman, M. (2009). Does correction for guessing reduce students' performance on multiple-choice examinations? Yes? No? Sometimes?. *Assessment & Evaluation in Higher Education, 34*(1), 1–15.

Bradbard, D. A., Parker, D. F., & Stone, G. L. (2004). An alternate multiple-choice scoring procedure in a macroeconomics course. *Decision Sciences Journal of Innovative Education, 2*(1), 11–26.

Budescu, D., & Bar-Hillel, M. (1993). To guess or not to guess: A decision-theoretic view of formula scoring. *Journal of Educational Measurement, 30*(4), 277–291.

Burton, R. F. (2001). Quantifying the effects of chance in multiple choice and true/false tests: Question selection and guessing of answers. *Assessment & Evaluation in Higher Education, 26*(1), 41–50.

Burton, R. F. (2002). Misinformation, partial knowledge and guessing in true/false tests. *Medical Education, 36*, 805–811.

Burton, R. F. (2004). Multiple choice and true/false tests: Reliability measures and some implications of negative marking. *Assessment & Evaluation in Higher Education, 29*(5), 585–595.

Bush, M. (1999). Alternative marking schemes for online multiple-choice tests. *7th annual conference on the teaching of computing.*

Bush, M. (2001). A multiple choice test that rewards partial knowledge. *Journal of Further and Higher Education, 25*(2), 157–163.
Choppin, B. H. (1988). Correction for guessing. In J. P. Keeves (Ed.), *Educational research, methodology, and measurement: An international handbook* (pp. 384–386). Oxford: Pergamon Press.
Coombs, C. H., Milholland, J. E., & Womer, F. B. (1955). The assessment of partial knowledge. *Educational and Psychological Measurement, 6*, 13–37.
Davis, F. B. (1967). A note on the correction for chance success. *Journal of Educational Measurement, 3*, 43–47.
Cohen-Schotanus, J., & van der Vleuten, C. P. M. (2010). A standard setting method with the best performing students as point of reference: Practical and affordable. *Medical Teacher, 36*, 154–160.
Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart & Winston.
de Gruijter, D. N. M. (2001). *Toetsing en toetsanalyse*. ICLON, Leiden: Sectie Onderwijsontwikkeling Universiteit Leiden.
Dochy, F., Kyndt, E., Baeten, M., Pottier, S., & Veestraeten, M. (2009). The effects of different standard setting methods and the composition of borderline groups: A study within a law curriculum. *Studies in Educational Evaluation, 35*, 174–182.
Downing, S. M. (2004). On guessing corrections. *Medical Education, 38*, 113.
Downing, S. M., Lieska, N. G., & Raible, M. D. (2003). Establishing passing standards for classroom achievement tests in medical education: A comparative study of four methods. *Academic Medicine, 78*(10), 85–87.
Dunn, L., Parry, S., & Morgan, C. (2007). *Seeking quality in criterion referenced assessment*.
Espinosa, M. P., & Gardeazabal, J. (2010). Optimal correction for guessing in multiple-choice tests. *Journal of Mathematical Psychology, 54*(5), 415–425.
Fowell, S., & Jolly, B. (2000). Combining marks, scores and grades. Reviewing common practices reveal some bad habits. *Medical Education, 34*, 785–786.
Frary, R. B. (1980). The effect of misinformation, partial information, and guessing on expected multiple-choice test item scores. *Applied Psychological Measurement, 4*, 79–90.
Frary, R. B. (1988). Formula scoring of multiple choice tests (correction for guessing). The entity from which ERIC acquires the content, including journal, organization, and conference names, or by means of online submission from the author. *Educational Measurement: Issues and Practice, 7*(2), 33–38.
Frary, R. B. (1989). Partial-credit scoring methods for multiple-choice tests. *Applied Measurement in Education, 2*(1), 79–96.
Gardner-Medwin, A. R. (1995). Confidence assessment in the teaching of basic science. *Research in Learning Technology, 3*(1), 80–85.
Hammond, E. J., McIndoe, A. K., Sansome, A. J., & Spargo, P. M. (1998). Multiple-choice examinations: Adopting an evidence-based approach to exam technique. *Anaesthesia, 53*, 1105–1108.
Jennings, S., & Bush, M. (2006). A comparison of conventional and liberal (free-choice) multiple-choice tests. *Practical Assessment, Research & Evaluation11(8)*.
Karandikar, R. L. (2010). On multiple choice tests and negative marking. *Current Science, 99*(8), 1042–1045.
Kubinger, K. D., Holocher-Ertl, S., Reif, M., Hohensinn, C., & Frebort, M. (2010). On minimizing guessing effects on multiple-choice items: Superiority of a two solutions and three distractors item format to a one solution and five distractors item format. *International Journal of Selection and Assessment, 18*(1), 111–115.
Kurz, T. B. (1999). A review of scoring algorithms for multiple-choice tests. *Paper Presented at the Annual Meeting of the Southwest Educational Research Association*.
Moss, E. (2001). Multiple-choice questions: Their value as an assessment tool. *Current Opinion in Anaesthesiology, 14*, 661–666.
Muijtjens, A. M. M., van Mameren, H., Hoogenboom, R. J. I., Evers, J. L. H., & van der Vleuten, C. P. M. (1999). The effect of a 'don't know' option on test scores: Number-right and formula scoring compared. *Medical Education, 33*, 267–275.
Ng, A. W. Y., & Chan, A. H. S. (2009). Different methods of multiple-choice test: Implications and design for further research. In *Proceedings of the international multiconference of engineers and computer scientists 20 (Vol. 2)*.
Norcini, J. J. (2003). Setting standards on educational tests. *Medical Education, 37*, 464–469.
Prieto, G., & Delgado, A. R. (1999). The effect of instructions on multiple-choice test scores. *European Journal of Psychological Assessment, 15*(2), 143–150.
Prihoda, T. J., Pinckard, R. N., McMahan, C. A., & Jones, A. C. (2006). Correcting for guessing increases validity in multiple-choice examinations in an oral and maxillofacial pathology course. *Journal of Dental Education, 70*(4), 378–386.
Rogers, H. J. (1999). Guessing in multiple choice tests. In G. N. Masters & J. P. Keeves (Eds.), *Advances in measurement in educational research and assessment* (pp. 235–243). Amsterdam: Pergamon.
Scharf, E. M., & Baldwin, L. P. (2007). Assessing multiple choice question (MCQ) tests – A mathematical perspective. *SAGE, 8*(1), 31–47.
Thayn, S. (2011). *An evaluation of multiple choice test questions deliberately designed to include multiple correct answers* Retrieved 28 June, 2011 from http://contentdm.lib.byu.edu/ETD/image/etd4168.pdf.
Traub, R. E., Hambleton, R. K., & Singh, B. (1969). Effects of promised reward and threatened penalty on performance of a multiple-choice vocabulary test. *Educational and Psychological Measurement, 29*, 847–861.
van der Vleuten, C. P. M. (2010). Setting and maintaining standards in multiple choice examinations: Guide supplement 37.1 – Viewpoint. *Medical Teacher, 32*, 174–176.
Zimmerman, D. W., & Williams, R. H. (2003). A new look at the influence of guessing on reliability of multiple-choice tests. *Applied Psychological Measurement, 27*(5), 357–371.