Evaluation & the Health Professions

<http://ehp.sagepub.com/>

Web-Based Application to Eliminate Five Contraindicated Multiple-Choice Question Practices

Andreas Brunnquell, Ümüt Degirmenci, Sebastian Kreil, Johannes Kornhuber
and Markus Weih

Eval Health Prof 2011 34: 226 originally published online 10 May 2010

DOI: 10.1177/0163278710370459

The online version of this article can be found at:

<http://ehp.sagepub.com/content/34/2/226>

Published by:

<http://www.sagepublications.com>

Additional services and information for *Evaluation & the Health Professions* can be found at:

Email Alerts: <http://ehp.sagepub.com/cgi/alerts>

Subscriptions: <http://ehp.sagepub.com/subscriptions>

Reprints: <http://www.sagepub.com/journalsReprints.nav>

Permissions: <http://www.sagepub.com/journalsPermissions.nav>

Citations: <http://ehp.sagepub.com/content/34/2/226.refs.html>

>> **Version of Record** - May 10, 2011

OnlineFirst Version of Record - May 10, 2010

What is This?

Web-Based Application to Eliminate Five Contraindicated Multiple-Choice Question Practices

Evaluation & the Health Professions

34(2) 226-238

© The Author(s) 2011

Reprints and permission:

sagepub.com/journalsPermissions.nav

DOI: 10.1177/0163278710370459

<http://ehp.sagepub.com>

**Andreas Brunnquell,¹ Ümüt Degirmenci,¹
Sebastian Kreil,¹ Johannes Kornhuber¹, and
Markus Weih¹**

Abstract

Multiple-choice questions (MCQs) evaluate factual knowledge in medical education and have a high reliability, if performed appropriately. However, many MCQs contain formal errors leading to reduced validity. The authors developed a Web application capable of recognizing and eliminating five frequent contraindicated practices in MCQs: negative stem, unfocused stem, cueing words, longest item = right item flaw, and stem/item similarities. The authors used simple string algorithms and dynamic comparisons with keywords. The system was successfully validated with a sample of approximately 800 continuous medical education (CME) questions, showing that our system automatically detects 60% of all formal didactic errors. Flaws

¹Department of Psychiatry and Psychotherapy, University Erlangen, GERMANY

Corresponding Author:

Markus Weih, Deptt. of Psychiatry and Psychotherapy, University Erlangen, GERMANY,
Schwabachanlage 6, 91054 Erlangen

Email: markus.weih@uk-erlangen.de

not detected by the software can easily be avoided using quick manuals on item wording or clear instruction to the authors. The authors conclude that it is feasible to improve the quality of MCQs by designing a Web application that is capable of detecting common flaws by simple string operations.

Keywords

medical education, assessment methods, test validity, undergraduate education, postgraduate education

Introduction

A common phrase states that students can escape bad teaching, but they cannot escape bad assessment. Assessment methods can be divided into “high-stake” tests (course graduation, promotion) and “low-stake” tests, mainly for directing student learning (Swanson, Clauser, & Case, 1999). Good assessment methods for “high-stake” tests have psychometric characteristics that are measurable, valid, objective, and closely related to the purpose of the test (Swanson, 1987). Multiple-choice questions (MCQs) are still the golden standard in most high-stake assessments of factual knowledge in undergraduate and postgraduate medical education. However, to reach sufficient validity, MCQs require standards like adequate difficulty, must be based on predefined learning objectives and should address real clinical problems. Therefore, good item writing is a difficult task, even for experienced authors.

Typical problems with older multiple-choice items are inadequate or contraindicated question formats like AOTA/NOTA (ALL-Of-The-Above and None-Of-The-Above). These formats are now not used any more in most instances because they decrease difficulty and validity of the test (Harasym, Leong, Violato, Brant, & Lorscheider, 1998). Examples of typical contraindicated practices in MCQ are given in Table 1 and the effects of contraindicated practices on test validity are given in Table 2.

Violation of standards in item writing decreases test validity via two mechanisms: (a) increase of construct-irrelevant test difficulty and variance (Downing, 2005) or (b) decrease of test difficulty when correct answers can be given by test wiseness of participants (“use the cues;” Case & Swanson, 2002; Krebs, R. Use the cues: Kurzttest zum erkennen ungewollter loesungshinweise bei multiple-choicefragen. personal communication, 30 March, 2004.) These problems occur independently from the language

Table 1. Typical Contraindicated Multiple-Choice Question (MCQ) Practices and Rationale Not to Use Them Any More

Contraindicated Practice	Effect on Assessment
Negative stems	Decreased readability; increased construct irrelevant test difficulty (Tamir, 1993)
Unfocused stem	Reduction of correct response rate (Downing, Baranowski, Grosso, & Powell, 1995)
Cues	Decreased test validity due to test wiseness (Downing, 2005)
Long answer = right answer flaw	Decreased test validity due to test wiseness (Case & Swanson, 2002)
Similarities between stem and option	Decreased test validity due to test wiseness (Case & Swanson, 2002)

used, because item writing is an inherent feature of the psychometric construct of the multiple-choice assessment format.

In the medical education literature, there are numerous guidelines that help authors in designing effective MCQs (Bloch et al., 1999; Case & Swanson, 2002; Smolle, 2009). In a previous study, we have shown that the quality of MCQs can be improved by simple peer-to-peer faculty development (Weih, 2008). For better illustration, Table 2 shows examples of common flaws in multiple-choice item writing.

The typical MCQ item author is a medical expert with time constraints, a lack of formal didactic education and state-of-the art item writing principles. As a consequence, formal errors in MCQ question stems and items are common. Stagnaro-Green et al. found flawed items in all of 40 *New England Journal of Medicine* CME questions (Stagnaro-Green & Downing, 2006). In a previous study, we found formal MCQ errors in 65% (518 of the 796) MCQs (Gutmann et al., 2009).

Kuehne-Eversmann, Nussbaum, Reincke, and Fischer (2007) found formal flaws in 68% of CME questions in German Medical Journals of Internal Medicine. This group also developed a taxonomy of 10 different types of formal flawed MCQ questions and items that we used in this article (Kuehne-Eversmann et al., 2007). Many of these formal flaws can in theory be detected by automated software pattern recognition. To our knowledge, this approach is innovative and we are not aware of any other computerized attempts in this area.

In the current study, we extended our peer-to-peer approach (Weih et al., 2008) and developed an automatic Web-based application to reduce flawed

Table 2. Examples of Common Contraindicated Practices in Item Writing

Contraindicated Practice	Example Of Typical Question Stem And Item
Negative stem	<p>Which statement about suicidal behaviour is <i>not</i> true? (Modified after Wolfersdorf, 2008)</p> <p>A: Ideas about committing suicide are an indicator for increased risk for completed suicide</p> <p>B: Passive thoughts about death are part of the suicidal spectrum</p> <p>C: Development of acute suicidality always follows a process of continuous development</p> <p>D: Acute suicidal impulses carry a high risk for completed suicide</p> <p>E: Patients talking about suicide have an increased risk</p>
Unfocused stem	<p>It is correct that (modified after Binkofski & Fink, 2005):</p> <p>A: In neglect there is always a head and gaze deviation towards the side with the lesion</p> <p>B: Neglect patients do react to sensory stimuli contralateral to the lesion side</p> <p>C: The lesion in neglect syndromes is within the primary sensory brain regions</p> <p>D: The angular gyrus in the right inferior parietal cortex is a critical region for neglect</p> <p>E: Remission of neglect is faster after lesions in the right hemisphere (vs. left)</p>

(continued)

Table 2 (continued)

Contraindicated Practice	Example Of Typical Question Stem And Item
Cues	<p>In patients with Alzheimer's dementia, memory deficit (modified after Case & Swanson, 2002)</p> <p>A: Can be treated with cholinesterase inhibitors</p> <p>B: <i>Could be due to deposition amyloid in the carotid artery</i></p> <p>C: <i>Is never seen in patients with neurofibril changes at autopsy</i></p> <p>D: <i>Is always severe</i></p> <p>E: <i>Possible involves the GABA-ergic system</i></p> <p>Secondary gain is</p> <p>A: <i>Synonym: Malingering</i></p> <p>B: <i>Frequent problem in compulsive disorder</i></p> <p>C: A complication of a variety of illnesses an tends to prolong many of them</p> <p>D: <i>Never seen in organic psychosis (Case & Swanson, 2002)</i></p> <p>A 58-year-old man with alcohol abuse is confused and agitated. He speaks of experiencing the world as unreal. This symptom is called:</p> <p>A: <i>depersonalisation</i></p> <p>B: <i>derailment</i></p> <p>C: derealisation</p> <p>D: <i>focal memory deficit (after Case & Swanson, 2002)</i></p> <p>Local anesthetics inactivate axons of neurons</p> <p>A: <i>In the anionic form, acting from inside the nerve membrane</i></p> <p>B: In the cationic form, acting from inside the nerve membrane</p> <p>C: <i>In the cationic form, acting from outside the nerve membrane</i></p> <p>D: <i>In the uncharged form, acting from inside the nerve membrane</i></p> <p>E: <i>In the uncharged form, acting from outside the nerve membrane (example after Case & Swanson, 2002)</i></p>
Longest answer = right answer flaw	
Similarities between stem and answer	
Word repetition in answer	

Note. Flaws, resp. Contraindicated practices are underlined and correct answers (resp. false in negative stems) are indicated in boldface.

MCQs in an item data bank. Currently, the system is programmed in German, but it can easily be transferred to other languages.

Method

First, a database of 796 CME questions was built from a common German neuropsychiatric journal as a starting point and to identify flaws by a didactic expert. Analysis of didactic errors was based on the Kuehne-Eversmann criteria (Kuehne-Eversmann et al., 2007). The analysis revealed that negative stems, unfocused stems, and cues were the most prevalent CME flaws. Then, with the help of a software programmer (see acknowledgment), there was an analysis on how to recognize CME flaws by recognition of PHP string operations and a discussion how to transform it to a Web-based application. Then, the suitable five most common didactic errors were transferred to regular expressions in an appropriately designed Web environment (PHP Version 5.2. for entry of user data and entry of MCQ formats. MySQL (Version 5.0) run on an Apache HTTP server (Version 2.2) with Linux as OS was used as a database system. The system was programmed with keywords, which can be entered and changed by an expert to reduce false negative or false positive results. The system allows checking new entered questions in a flexible manner. Examples of PHP scripts with coding algorithms for detection of negative worded stems; entry of keywords and for detection of similarities using the Levensthein-Distance are given in Figures 1–3.

The software implementation was accomplished as follows: The negative question (5), item and cueing (9.1) flaws are recognized by the database by allowing the administrator to enter particular predefined keywords into a string repository (negative flaw words were 5: “non, not;” cueing words were 9.1: “always, never, usually, only, alone etc.” as indicated by Case & Swanson, 2002). After entry, the system then allows dynamic direct comparison of these keywords with the new MCQ items and stems entered by the question author.

Unfocused stem (4) was defined as a question stem that falls below a predefined threshold representing a brief question stem. This threshold can be predefined by the expert. Our analysis of CME questions with flaws showed that an unfocused stem usually has string length of 21 or lower (e.g., “Which answer is right?” = string length with spaces = 21). The longest = right answer flaw (9.4) is implemented by calculating the mean item string length. Then, the system calculates if a single item fulfills the following logical conditions: length exceeding > 10 (predefinable) values

```

    $SQL_KEYWORDS="SELECT ID, info, attributes FROM questions_info
    WHERE ID=1"; // ID=1 is for questions key words!

    $rs = $DB_conn->Execute($SQL_KEYWORDS);

    $SQL='SELECT ID, question FROM questions WHERE (';

    foreach($keyword as $keyw)

        return '<font color="red">[Flaw] Negative Question Stem! Should be
        used with caution !.</font><br />'; else

        return '<font color="green">OK. No negations.</font><br />'; }

```

Figure 1. PHP-SQL-pseudo algorithm to detect negative multiple-choice questions (MCQ) stems. Note: The complete PHP code is available from the authors upon request.

```

mysql> select * from questions_info ego where id=1\G

    ID: 1 info: Negative question stem: All question stems with the following
attributes:

attributes: no, not

expert: SELECT ID, Questions from QUESTIONS WHERE questions LIKE

"%no%" OR questions LIKE "%not%"

```

Figure 2. PHP-SQL-pseudo algorithm for cueing keywords. Note: The complete PHP code is available from the authors upon request.

above mean AND right item = YES. Similarities in between items (9.5) and between items and stems are implemented into the system by calculating the Levenshtein-Distance between the strings. A distance of 0 indicates a string repetition (for details, the complete algorithm is displayed in Figure 3).

Results

The following formal MCQ flaws could be successfully tested in the Web-based application and tested according to the taxonomy of


```

function do_plausi_a_similar($DB_conn,$QID)
{
    $SQL='SELECT ID, answer FROM answer WHERE fk_question='.$QID;

    $rs = $DB_conn->GetArray($SQL);

    // init a rray for levenshtein distance

    $lev_dist = array(); // outer loop

    foreach($stack as $key_o => $value_o){

        // Levenshtein distance between items

        array_push($lev_dist,

        levenshtein(substr($value_o,0,255), substr($value_i,0,255)));} } /*

    // word count

    foreach($ stack as $key => $value) {

        $word_arr = count_chars($value,1);

        print_r($word_arr) print "<br>";} print "<br>"; /*

        $ret = '<font color="green">Levenshtein:

        '.min($lev_dist).'

```

Figure 3. PHP-SQL-pseudo algorithm to detect Levenshtein similarities between question stem and items. Note: The complete PHP code is available from the authors upon request.

Kuehne-Eversmann and Fischer (numbers in brackets; Kuehne-Eversmann et al., 2007):

- (5) negative question stem or item
- (4) unfocused stem
- (9.1) cues within items
- (9.4) longest item = right item flaw
- (9.5–9.7) similarities or word repetition between items or between stem and item.

Figure 4. Entry of multiple-choice questions (MCQ). Notes: Data term entry system of MCQ for question authors. MCQ stem is left and items are on the right side. The correct item is indicated by a tick-box. The button “Generate Test” edits a complete test with a selection of MCQs. The button “Check Questions” controls for common flaws as indicated in the text. “Import” allows usage of questions and items arranged by a spreadsheet program as common separated value (CSV). Access to the system is available upon request.

After implementation of common didactic flaws in MCQs to regular expressions using PHP-scripts (Figure 1; as described in Method), the system was tested. Figure 2 shows a screenshot of the data entry system.

The test database of 796 CME-MCQs showed 647 item flaws. The detailed distribution of the flaws is given in our previous article (Gutmann, Degirmenci, Kornhuber, & Weih, 2010) or is available as a table upon request to the authors. Figure 4 shows the data entry mask, and Figure 5 the recognition of typical cues in the MCQ stem. To validate the system in a larger setting, an MCQ database with 796 questions and manually identified item flaws (as described in Method) was tested with the system. We have shown before that the distribution of formal flaws in this data set is as follows

- (5) negative question stem, negative item or double negation flaw: 181/643 = 28.1%
- (4) unfocused stem flaw: 47/643 = 7.3%
- (9.1) cueing (e.g., always, never, usually, only, alone, and exclusively): 112/643 = 17.4%
- (9.4) longest answer = right answer flaw: 30/643 = 4.7%
- (9.5) similar item answers flaw: 16/643 = 2.5%
- (9.6) verbal similarities/associations between stem and answer flaw: 2/643 = 0.3%

<p>In the CLINICAL ASSESSMENT OF CHRONIC PAIN ...</p> <p>A</p>	<ol style="list-style-type: none"> 1. ☐ UNPLEASED EMOTIONS MAY BE CONVERTED TO COMPLAINTS OF BOOPLY PAIN 2. ☒ THE PHYSICIANS PERSONAL ATTITUDE CONCERNING PAIN MAY AFFECT MEDICAL JUDGEMENT 3. ☐ PAIN MAY HAVE A SYMBOLIC MEANING 4. ☐ FACIAL APPEARANCE OF BODY POSTURE IS USUALLY A CLUE TO THE SEVERITY OF PAIN <p>Error! One or more cues found in items!</p>
<p>WHICH ANSWER BEST DESCRIBES THE CLINICAL ASSESSMENT OF CHRONIC PAIN ...</p> <p>B</p>	<ol style="list-style-type: none"> 1. ☐ UNPLEASED EMOTIONS ARE RARELY CONVERTED TO COMPLAINTS OF BOOPLY PAIN 2. ☒ THE PHYSICIANS PERSONAL ATTITUDE CONCERNING PAIN AFFECTS MEDICAL JUDGEMENT 3. ☐ PAIN HAS A SYMBOLIC MEANING IN WESTERN CULTURE 4. ☐ FACIAL APPEARANCE OF BODY POSTURE IS A CLUE TO THE SEVERITY OF PAIN <p>OK! No cue in item.</p>

Figure 5. Examples for items before and after automatic feedback given by the software. Note: Left side: MCQ with the cues “may” or “usually” before (A) with red error flag and after (B) flaw recognition (Green OK flag). Right side: MCQ with items of different length. The longest item is the right answer = longest answer flaw (C). After revision, the item is flagged green (D)

After adaptation of the index words, the false negative and false positive rate of the system was reduced to approximately 5%. With a keyword repository of 16 words for negative stems and cues, the system is now able to identify all five flaws as described in the introduction, which together comprise 60% of contraindicated flaws. Access to the test system is available under develop.netz.org/login.php.

Discussion

As a main result of the current study, we are able to show that common formal flaws in MCQs can easily be transferred to regular expressions in an appropriately designed IT system. Our automatic software application is currently able to recognize five common flaws that comprise approximately 60% of all item flaws in our test database of MCQs. In our test database, approximately 30% of the other flaws consisted of the AOTA or NOTA-type (All resp. None-Of-The-Above item type) or item combinations (e.g., five options 1–5 and A–E give various combinations of 1–5). These flaws can easily be avoided by appropriate instruction of the question authors or by reading a quick manual on item wording like Case and Swanson, 2002. Currently, our system gives feedback in German language, but the software algorithms can easily be transferred to other languages because the inherent

problems of multiple-choice item writing constructs are independent of the language used.

Assessment is a crucial part in each phase of a medical professionals training. Assessment results provide important information about a learner's proficiency and progress of learning during education. The score in a high-stake assessment can be the most important indicator of the extent of knowledge in the tested domain. However, any assessment is only a small sample of the real knowledge in the tested domain, because real-life assessments are always limited. Measurement is termed valid when a test score is assumed to reflect real performance. Multiple-choice assessments are the most prevalent testing instrument for knowledge in medical education and are based on the classical test theory. This theory assumes that a given test candidate, on a given test (X), has a true score (T) and a random error component (E):

$$X = T + E. \quad (1)$$

In reality, multiple-choice tests deviate from the test theory due to several reasons. First, the random component always influences the true score. For example, a difficult 5-item question allows 20% random correct numbers. Then, experienced examinees develop test wiseness, which allows higher scores due to logical operations alone. On the examiners' side, the main workload in MCQs is the adequate preparation of stems and items, which requires both expert and didactic knowledge. As a rule of thumb, the resources required for one question are about 1 hr of expert's time. Because valid MCQ should have 100 questions or more, the preparation time for a valid test is substantial. In reality, time restrictions and a lack of formal training of experts decreases the quality of many multiple-choice tests. Accordingly, we and other groups have shown that a large percentage of MCQs contains formal errors (Kuehne-Eversmann et al., 2007; Stagnaro-Green & Downing, 2006).

In theory, many formal errors can be categorized and detected by MCQ software. There are an increasing number of commercially available inexpensive software packages on the English-speaking market such as BILOG-MG, WINSTEPS, FACETS, PARSQUALE, or HOT POTATOES (reviewed by De Champlain, 2010). In Germany, ITEM MANAGEMENT SYSTEM (www.ims-m.de; accessed February 15, 2010) and FRED (www.blubbsoft.de; accessed February 15, 2010) are used on the free market. The German Medical Assessment Institute (IMPP) has opened an MCQ database for commercial use (SPIDMED: www.impp.de; accessed February 15,

2010). However, to our knowledge, none of these software solutions and data banks for MCQs provides intelligent recognition of common formal question flaws.

In summary, we were able to develop a Web-based multiple-question database with recognition of common contraindicated practice. Automatic analysis of five common MCQ flaws was successfully programmed, implemented, and tested in our system and controlled against a large database of approximately 800 MCQs from a CME journal. These flaws represent a large portion of common formal errors in MCQs. Our system avoids time-consuming manual correction by a didactic trained peer or expert. This could help in reducing common didactic flaws and could thereby be a valuable tool to increase the quality of MCQs tests.

Acknowledgment

The authors would like to thank S. Maeuser, D. Ettle (IT technicians) and J. Hercher, CEO of Mobited Corp., Germany (www.mobited.de) for software development support.

Declaration of Conflicting Interests

The author(s) declared no conflicts of interest with respect to the authorship and/or publication of this article.

Funding

The author(s) received no financial support for the research and/or authorship of this article.

References

- Binkofski, F., & Fink, G. (2005). Apraxias. *Nervenarzt*, 76, 493-509, quiz 510-491.
- Bloch, R., Hofer, D., Krebs, R., Schlaeaeppi, P., Weiss, S., & Westkaeemper, R. (1999). *Kompetent pruefen: Handbuch zur Planung, Durchfuehrung und Auswertung von Facharztpruefungen*. University of Bern, Switzerland: University Press.
- Case, S. M., & Swanson, D. B. (2002). *Constructing written test questions for the basic and clinical sciences* (3rd Rev. ed.). Philadelphia, PA: National Board of Medical Examiners (NBME).
- De Champlain, A. F. (2010). A primer on classical test theory and item response theory for assessments in medical education. *Medical Education*, 44, 109-117.
- Downing, S. M. (2005). 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. *Advances in Health Sciences Education: Theory and Practice*, 10, 133-143.
- Downing, S. M., Baranowski, R. A., Grosso, L. J., & Powell, R. D. (1995). Item type and cognitive ability measured: The validity evidence for multiple true-false item in medical speciality certification. *Applied Measurement in Education*, 8, 89-99.
- Gutmann, A., Degirmenci, U. E., Kornhuber, J., & Weih, M. (in press). Formale analyse der CME-Fragen aus "Der Nervenarzt" von 2002-2008. *Der Nervenarzt*, 8.
- Harasym, P. H., Leong, E. J., Violato, C., Brant, R., & Lorscheider, F. L. (1998). Cuing effect of "all of the above" on the reliability and validity of multiple-choice test items. *Evaluation & the Health Professions*, 21, 120-133.
- Kuehne-Eversmann, L., Nussbaum, C., Reincke, M., & Fischer, M. R. (2007). CME activities of medical journals: Quality of multiple-choice questions as evaluation tool. Using the example of the German medical journals *Deutsches Arzteblatt*, *Deutsche Medizinische Wochenschrift*, and *Der Internist*. *Medizinische Klinik (Munich)*, 102, 993-1001.
- Smolle, J. (2009). *Klinische MC-Fragen rasch und einfach erstellen*. de Gruyter, Berlin, New York.
- Stagnaro-Green, A. S., & Downing, S. M. (2006). Use of flawed multiple-choice items by the New England Journal of Medicine for continuing medical education. *Medical Teacher*, 28, 566-568.
- Swanson, D. B. (1987). A measurement framework for performance-based tests. In I. Hart & R. Harden (Eds.), *Further developments in assessing clinical competence*. Montreal, Canada: Can-Heal Publications.
- Swanson, D. B., Clauser, B. E., & Case, S. M. (1999). Clinical skills assessment with standardized patients in high-stakes tests: A framework for thinking about score precision, equating, and security. *Advances in Health Sciences Education: Theory and Practice*, 4, 67-106.
- Tamir, P. (1993). Positive and negative multiple-choice items: How difficult are they? *Studies in Educational Evaluation*, 19, 311-332.
- Weih, M., Harms, D., Rauch, C., Segarra, L., Reulbach, U., Degirmenci, U., & ... Kornhuber, J. (2008). Qualitätsverbesserung von multiple-choice Prüfungen in psychiatrie, psychosomatik, psychotherapie und neurologie. *Der Nervenarzt*, 80, 324-328.
- Wolfersdorf, M. (2008). Suicidalität. *Nervenarzt*, 79, 1319-1334;quiz 1335-1316.