Effects of automatic item eliminations based on item test analysis

Jaap H. J. Muntinga and Henk A. Schuil

Department of Medical Physiology, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands

Submitted 29 March 2007; accepted in final form 31 May 2007

Muntinga JH, Schuil HA. Effects of automatic item eliminations based on item test analysis. Adv Physiol Educ 31: 247–252, 2007; doi:10.1152/advan.00019.2007.—Item test analysis is an aid to identify items that need to be eliminated from an assessment. An automatic elimination procedure based on item statistics, therefore, could help to increase the quality of a test in an objective manner. This was investigated by studying the effect of a standardized elimination procedure on the test results of a second-year course over a period of 6 successive years in 1,624 candidates. Cohort effects on the item elimination were examined by determining the number of additional items that had to be eliminated from three different tests in 3 successive academic years in two cohorts. The items that were part of more than one test and had to be eliminated according to the procedure in at least one of the tests appeared to have to be retained according to the same procedure in most of the other tests. The procedure harmed the high scoring students relatively more often than the other students, and the number of eliminated items appeared to be cohort dependent. As a consequence, automatic elimination procedures obscure the transparency of the grading process unacceptably and transform valid tests into inadequate samples of the course content.

METHODS

Participants

The elimination procedure was investigated by studying the effects of the procedure on the test results of a second-year course in medicine on renal physiology. From the academic years of 1998/1999 through 2003/2004, a total of 1,624 candidates were tested in this course spread over six regular multiple-choice tests. Furthermore, the number of items eliminated by the procedure was determined in the same cohort across three different year levels. It was calculated for the cohort starting in 2000 with 332 participants in the selected first-year test and 288 participants in the selected third-year test and for the cohort starting in 2001 with 330 participants in the selected first-year test and 288 participants in the selected third-year test.

Analysis

Multiple-choice questions were analyzed by calculating the item difficulty index (P value) and the item total correlation (Rit score). The P value is the number of candidates who answered the item correctly divided by the total number of students who answered the item. The lower the P value, the more difficult a particular item is. The Rit score is the point-biserial correlation between the item score and test score. A high correlation means that candidates who did well on the item also did well on the test. A negative correlation means that candidates who performed well on the test did not choose the correct item response. The Rit score consequently shows the degree to which each item discriminates between candidates with high scores and those with low scores and therefore represents a discrimination index.

In the traditional subprocedure, a student delegation of the year class makes an inventory of all student feedback. Additionally, items with a P value of <0.30 and/or a Rit score of <0.20 in the item test analysis are flagged as suspicious. Items with valid student feedback and items marked as suspicious are inspected subsequently for grammatical errors, inconsistencies, obvi-

Address for reprint requests and other correspondence: J. H. J. Muntinga, Dept. of Medical Physiology, Univ. Medical Center Groningen, A Deusinglaan 1, Groningen 9713 AV, The Netherlands (e-mail: j.h.j.muntinga@med.umcg.nl).
How We Teach

AUTOMATIC ITEM ELIMINATION

Table 1. Classification of items in 15 classes according to P values and Rit scores used by the item elimination procedure under investigation

<table>
<thead>
<tr>
<th>Class</th>
<th>P Value of an Item</th>
<th>Rit Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>With 2 Alternatives</td>
<td>With 3 Alternatives</td>
</tr>
<tr>
<td>0.92 &lt; P</td>
<td>0.90 &lt; P</td>
<td>0.89 &lt; P</td>
</tr>
<tr>
<td>0.77 &lt; P ≤ 0.92</td>
<td>0.70 &lt; P ≤ 0.90</td>
<td>0.66 &lt; P ≤ 0.89</td>
</tr>
<tr>
<td>0.63 &lt; P ≤ 0.77</td>
<td>0.50 &lt; P ≤ 0.70</td>
<td>0.44 &lt; P ≤ 0.66</td>
</tr>
<tr>
<td>0.50 &lt; P ≤ 0.63</td>
<td>0.33 &lt; P ≤ 0.50</td>
<td>0.25 &lt; P ≤ 0.44</td>
</tr>
<tr>
<td>P ≤ 0.50</td>
<td>P ≤ 0.33</td>
<td>P ≤ 0.25</td>
</tr>
</tbody>
</table>

The P<value> value is the P<value> value corrected for guessing, and Rit<value>min is the lowest Rit<value> score significant on the 5% significant level. E<value> type classes represent items that have to be eliminated according to the procedure. R<value> type classes represent items that have to be retained, and the P<value> type class represents suspicious items that may be eliminated after a second item inspection.

Grading

After each subprocedure, all candidates were graded using the following formula: test score = (score – guess score) × 10/(maximum score – guess score), where “score” is the number of items correctly answered by the candidate; “maximum score” is the score if all items of the test would have been answered correctly; and “guess score” is the score if all items of the test would have been answered by random guessing alone.

If an item has k alternatives, the probability of a correct guess amounts to 1/k. By multiplying this probability with the number of items with k alternatives, the guess score of all items with k alternatives was computed. The total guess score of all items of the test together was calculated by adding all guess scores of items with a different number of alternatives. A student failed for the test if the test score was lower than 5.5.

Test Analysis

Finally, the reliability and difficulty index of the test were determined. The reliability of the test was investigated using the Kuder-Richardson 20 (KR20) formula. The arithmetic average of the scores was used as the difficulty index of the test.

Statistics

All item test analyses were performed independently of our department by an institute specialized in the statistically evaluation of test results: the assessment service of the University Educational Centre. Paired comparisons were performed using the Wilcoxon signed-ranks test and unpaired comparisons with the Mann-Whitney test. P values of <0.05 were considered statistically significant. All statistical analyses were performed using a commercial software package (SPSS version 12.00 for Windows, SPSS).

RESULTS

Table 2 shows the main effects of the three subprocedures on the regular multiple-choice tests of the second-year course. The six tests refer to the same course with respect to design and teachers as well as learning objectives but were given in 6 different successive years. In these six tests, 30 items were eliminated by the traditional subprocedure.

False/True Items

All false/true items in the tests were simple, not multiple false/true items. After following the traditional subprocedures for all 6 tests, a total of 567 items remained to test 1,624 candidates. Of these items, 297 items (52.38%) were false/true items.

Due to the minimum subprocedure, a total of 32 extra items was eliminated. Most of these items (65.63%) were classified in class 15 (10 items) and class 10 (11 items) of the automatic item elimination procedure; 21 of the eliminated items (65.63%) were false/true items.

Due to the maximum subprocedure, a total of 18 items was eliminated additionally. Of these items, 12 items (66.67%) were false/true items.

Recurrent Items

Some of the items deleted during the minimum or maximum subprocedure were part of more than one test: 15 different multiple-choice questions were found in more than one test. Figure 1 shows how often each of these questions was classified in which main class according to the automatic item elimination procedure.

Effect on Grading

From failed to passed. After following the traditional subprocedure, 71.77 ± 12.08% of the 270.7 ± 41.1 students of a year class passed the exam (Table 2). After following the minimum procedure, 78.21 ± 8.95% of the students of a year class passed the exam, and after following the maximum procedure, 78.39 ± 11.76% of the students of a year class passed the exam. In other words, 6.44% of the students of a year class due to the minimum subprocedure and 6.62% of the students of a year class due to the maximum procedure were passed additionally.

Advances in Physiology Education • VOL 31 • SEPTEMBER 2007

Downloaded from http://advan.physiology.org/ by 10.220.33.4 on June 22, 2017
Table 2. Effects of 3 different item elimination procedures on the number of remaining test items, reliability of the test, difficulty of the test, and percentage of passed candidates of a year class

<table>
<thead>
<tr>
<th>Test Year</th>
<th>1998</th>
<th>1999</th>
<th>2000</th>
<th>2001</th>
<th>2002</th>
<th>2003</th>
<th>Means ± SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Candidates</td>
<td>233</td>
<td>246</td>
<td>257</td>
<td>253</td>
<td>290</td>
<td>345</td>
<td>270.7 ± 41.1</td>
</tr>
<tr>
<td>Number of test items</td>
<td>77</td>
<td>92</td>
<td>89</td>
<td>117</td>
<td>118</td>
<td>94</td>
<td>97.8 ± 16.3</td>
</tr>
<tr>
<td>Number of test items</td>
<td>After SPrad</td>
<td>75</td>
<td>87</td>
<td>87</td>
<td>114</td>
<td>113</td>
<td>91</td>
</tr>
<tr>
<td>After SPmin</td>
<td>70 (6.7)</td>
<td>83 (4.6)</td>
<td>79 (9.2)</td>
<td>104 (8.8)</td>
<td>110 (2.7)</td>
<td>89 (2.2)</td>
<td>89.2 ± 15.3 (5.7 ± 3.0)</td>
</tr>
<tr>
<td>After SPmax</td>
<td>68 (9.3)</td>
<td>81 (6.9)</td>
<td>74 (14.9)</td>
<td>98 (14.0)</td>
<td>109 (3.5)</td>
<td>87 (4.4)</td>
<td>86.2 ± 15.3 (10.5 ± 4.7)</td>
</tr>
<tr>
<td>Percentage of passed candidates</td>
<td>After SPrad</td>
<td>71.25</td>
<td>73.17</td>
<td>70.43</td>
<td>50.59</td>
<td>77.93</td>
<td>87.25</td>
</tr>
<tr>
<td>After SPmin</td>
<td>79.83</td>
<td>72.76</td>
<td>83.66</td>
<td>63.24</td>
<td>81.38</td>
<td>88.41</td>
<td>78.21 ± 8.95</td>
</tr>
<tr>
<td>After SPmax</td>
<td>81.55</td>
<td>76.42</td>
<td>85.99</td>
<td>55.73</td>
<td>83.10</td>
<td>87.54</td>
<td>78.39 ± 11.76</td>
</tr>
<tr>
<td>KR20</td>
<td>After SPrad</td>
<td>0.7293</td>
<td>0.8222</td>
<td>0.7867</td>
<td>0.7750</td>
<td>0.7412</td>
<td>0.7577</td>
</tr>
<tr>
<td>After SPmin</td>
<td>0.7928</td>
<td>0.8272</td>
<td>0.8008</td>
<td>0.7801</td>
<td>0.8580</td>
<td>0.7638</td>
<td>0.8038 ± 0.0340</td>
</tr>
<tr>
<td>After SPmax</td>
<td>0.7994</td>
<td>0.8322</td>
<td>0.8200</td>
<td>0.7929</td>
<td>0.8598</td>
<td>0.7733</td>
<td>0.8129 ± 0.0309</td>
</tr>
<tr>
<td>Test difficulty</td>
<td>After SPrad</td>
<td>0.7989</td>
<td>0.7703</td>
<td>0.7554</td>
<td>0.7152</td>
<td>0.7638</td>
<td>0.8102</td>
</tr>
<tr>
<td>After SPmin</td>
<td>0.7918</td>
<td>0.7810</td>
<td>0.7970</td>
<td>0.7424</td>
<td>0.8047</td>
<td>0.8174</td>
<td>0.7891 ± 0.0259</td>
</tr>
<tr>
<td>After SPmax</td>
<td>0.7949</td>
<td>0.7874</td>
<td>0.8059</td>
<td>0.7478</td>
<td>0.8079</td>
<td>0.8246</td>
<td>0.7948 ± 0.0263</td>
</tr>
</tbody>
</table>

Lower marks. At the same time, 7.53 ± 8.70% of the students of a year class due to the minimum subprocedure and 8.60 ± 7.23% of the students of a year class due to the maximum procedure received a lower mark. Summarized over all 6 years, 137 of the 1,624 candidates (8.44%) received a lower mark due to the minimum subprocedure and 139 of the 1,624 candidates (8.56%) received a lower mark due to the maximum subprocedure. The candidates who were lower graded by the minimum subprocedure scored 6.45 ± 1.96 after performing the traditional subprocedure and 6.38 ± 1.98 after performing the minimum subprocedure. The other candidates who were not lower graded scored 6.30 ± 1.43 and 6.66 ± 1.44, respectively. The distribution of the test scores is shown in Fig. 2. The test scores of the students who were lower graded by the minimum subprocedure decreased significantly with 0.064 ± 0.042 but increased with 0.363 ± 0.250 in other students. The candidates who were lower graded after performing the maximum subprocedure scored 5.43 ± 1.82 according to the traditional subprocedure and 5.31 ± 1.85 according to the maximum subprocedure. The other candidates scored 6.39 ± 1.42 and 6.80 ± 1.46, respectively. The test scores of the students who were lower graded by the maximum subprocedure decreased significantly with 0.124 ± 0.108 but increased with 0.414 ± 0.304 in other students.

DISCUSSION

Neither the criteria of the item elimination procedure nor the grading using a correction for guessing were the subjects of investigation in this study. In this study, we investigated the effects of an automatic elimination procedure on some important characteristics of a test such as its capability to determine what and how much had been learned by the students, the number of passed students, and the discrimination between candidates.

The items eliminated by both the minimum and the maximum procedures, and hence according to the automatic item elimination procedure, referred relatively more often to multiple-choice questions of the false/true type than to multiple-choice questions of other types. This agrees with the earlier findings of others demonstrating that false/true items are more susceptible to discussion (2, 5).

Are the Students or Items Bad?

Most of the items eliminated by the minimum procedure appeared to be classified in classes 10 and 15. The most decisive role in the forced elimination of items during the automatic elimination procedure, consequently, is played by a P value lower than the guess score. Such a low P value may
indicate that the item was flawed, but it may also indicate that the item was too difficult for most of the candidates. Because flawed items should be eliminated from an exam but extremely difficult (and easy) items are needed to adequately sample course content and objectives (13, 19), more insight in the cause of the low $P$ values is crucial. Since the Rit score strongly decreases not only at high $P$ values but also at low $P$ values by definition, the Rit score is of less importance in solving this problem.

If the items eliminated by the automatic elimination procedure were bad instead of difficult, than the students who got a lower test score due to the elimination procedure answered these items correctly only by chance. As a consequence, one would expect that the test scores of these students were at best distributed in the same way as those of the other students, but at least not higher. The students who got a lower test score due to automatic item elimination according to the minimum subprocedure, however, scored above average compared with other students (Fig. 2). The percentage of students who got a lower test mark due to automatic item elimination procedure amounted to 8.4% of all students, but appeared to be increased in the best (as well as the worst) performing quartile of the students (Fig. 3). In the 10% highest scoring students, for example, this percentage even amounted to 19% (31 students). As a consequence, the automatic item elimination harmed the higher scoring students more often than other students. This suggests that many of the eliminated items are not bad but merely difficult or refer to objectives that were not learned by most of the students.

The question of whether it is justified to use low $P$ values and Rit scores as indisputable arguments in a largely automatic item elimination procedure was investigated further by comparing the $P$ values and Rit scores of the eliminated items that appeared in more than one test. The large variation in the classification of these items over the various tests (Fig. 1) convincingly illustrates that the $P$ value and Rit score of a multiple-choice question not only depend on the item itself but on the test population and test circumstances as well. As demonstrated previously (7), the item test statistics were hopelessly confounded with the particular sample of examinees who completed the assessment. Most of the recurrent items that had to be eliminated (class $E$) according to the automatic item elimination procedure in one of the tests could (class $PE$) or even had to be retained (class $R$) according to the same procedure in most of the other tests. This finding illustrates in another way but again that most of the eliminated items were eliminated not because they were flawed but because they were difficult or refer to course objectives that were not learned by most of the students.

The question of whether the items or the students are bad turns up once more if the relatively large number of eliminated items is noticed in the assessment of year 2001 compared with that of year 2002 (Table 2). This large number could have been caused by poor quality of the items, a large number of difficult items, inferior teaching, or even by factors related to the students themselves as well. Therefore, for the students who were tested for the second-year course in 2001 (cohort 2000), we determined the number of additional items that had to be
eliminated unconditionally by the automatic elimination procedure (i.e., according to the minimum procedure) compared with the traditional procedure in both a first-year course and third-year course. Similar numbers were determined for the students who were tested for the second-year course in 2002 (cohort 2001). Figure 4 shows the resulting numbers of additionally eliminated items as a percentage of the total numbers of test items for both cohorts. These percentages were significantly higher for all three tests in cohort 2000 compared with cohort 2001. Because these tests refer to three completely different courses with different learning objectives and different teachers in different academic years, it is rather improbable that the teaching was significantly worse, the quality of the test items significantly lower, or the item difficulty significantly higher in all three courses of one specific cohort. The results shown in Fig. 4 suggest again, therefore, that many of the items eliminated by the automatic elimination procedure were eliminated not because they were bad but because of factors related to the student cohort.

Because most of the recurrent items that had to be eliminated in one test had to be retained in other tests, because the elimination procedure harmed the high-scoring students more often than other students, and because the number of eliminated items appeared to be cohort dependent, there is little reason to believe that these items are flawed and missed by the thorough postassessment inspection described above.

Effects of the Elimination Procedure

The number of passed students was significantly increased by the minimum subprocedure. By eliminating the items of class $PE$ as well, the number of passed students was hardly changed. That is, if all items that may be eliminated according to the automatic item elimination procedure were flawed items, the elimination of class $PE$ items in addition to the forced elimination of class $E$ items adds little to the decision of whether a student has passed or failed the exam. From this point of view, the elimination procedure could be simplified by restricting the item classification to two classes (classes $E$ and $R$) instead of three classes.

If the items were eliminated not because they were flawed but because of factors related to the students, as suggested above, the elimination procedure reduces to an ordinary relative grading process. The formula that was used for calculating the test score, however, suggests that the students were graded using an absolute instead of a relative grading procedure. As a consequence, the automatic elimination procedure obscures the transparency of the grading process by converting an apparently absolute grading process into a relative one.

As frequently demonstrated (3, 6, 9), exams have a decisive impact on the study behavior of students: “The impact of assessment on the educational process is a variable which allows little compromise” (17), “The assessment tail wags the curriculum dog” or “Grab the students by the tests and their hearts and minds will follow” (16), and “Students don’t learn what you expect, they learn what you inspect” (11). Because difficult and unpopular items are answered correctly by only a very small percentage of students, by definition, such items receive low $P$ values in test analyses and will be eliminated irrevocably during automatic elimination procedures. Each time a student is tested, therefore, the student will learn that s/he is not tested on difficult or unpopular learning objectives. The repeated use of the automatic item elimination procedure, in other words,
creates a hidden curriculum (10) in which our future doctors learn to avoid the more difficult and unpopular problems.

If items eliminated by the automatic elimination procedure are not eliminated because they are flawed, the procedure reduces a test to an inadequate sample of the course objectives. Hence, the validity of a test is diminished by the elimination procedure, although the test statistics in numbers seem to be improved (Table 2). In this study, consequently, arguments were found to justify the concerns reported earlier by others: “not to eliminate all items from a test which show poor item statistics regardless of their content” (15).

Comparison With Other Elimination Procedures

Compared with criteria of the investigated automatic item elimination procedure, more established criteria for malfunctioning items usually use higher values for the lower limit of an acceptable Rit score (8, 14) as well as P value (1). Fully automatic item elimination procedures according to such criteria, therefore, will lead to even more inequitable item eliminations.

The use of $R_{it, min}$ as the minimum value of acceptable Rit scores in the used classification differs from that of most other classifications. Usually a fixed Rit score is used as the minimum value (8, 14). $R_{it, min}$ varies inverse with the number of candidates. Because of the large numbers of students in medicine, the $R_{it, min}$ is generally lower than minimum values used by other classifications. Hence, fewer items will be eliminated on account of a low Rit score by the procedure. An item with a Rit score just below $R_{it, min}$ on the other hand, has to be eliminated from a test with little candidates but may be retained in a test with many candidates. According to the automatic item elimination procedure, therefore, one and the same item may lower the test quality unacceptably in one test but not in the other, depending on the numbers of candidates solely.

Since the quality of an item has nothing to do with the numbers of candidates, the use of a fixed limit for the Rit score has to be preferred in item test analyses.

Alternatives to the automatic item elimination procedure, such as item response theory (IRT) measurement models, were not investigated in this study. One of the main advantages of IRT measurements over classical test analysis, however, appears to be that it is not sample dependent, like the item elimination procedure we investigated in this study, and that it dissects apart student proficiency and the inherent difficulty of individual test items (7). IRT, on the other hand, requires data from many hundreds of individuals and seems to be less appropriate to the deliberate miscellaneous of knowledge and ability that are typically assessed in medical examinations. It is indeterminate, therefore, which of the methods is obviously better than the other (4). No matter how, using either classical test analysis or one of the alternative methods, the evaluation of the quality of a test by statistical analyses alone seems to be inadequate.

Conclusions

Item test analysis does help identify items that may be considered for elimination after postassessment inspection by a student delegation and responsible teachers. The quality of a test, however, appears not to be increased by an automatic elimination procedure since most of the eliminated items are not flawed at all. Automatic item elimination obscures the transparency of the grading process, is often harmful to high performing students, encourages future doctors to avoid difficult and unpopular problems, and transforms valid tests into inadequate samples of the course objectives. The evaluation of the quality of a test, consequently, has to be a mix of statistical analyses and, above all, human judgment, in which content review by students and peers is imperative.

ACKNOWLEDGMENTS

The authors acknowledge all members of the assessment service of the University Educational Centre of Groningen for the constructive participation in item test analyses.

REFERENCES