Certification of Airport Security Officers Using Multiple-Choice Tests: A Pilot Study

Diana Hardmeier  
Center for Adaptive Security Research and Applications (CASRA)  
Zurich, Switzerland

Catharina Müller  
School of Applied Psychology  
University of Applied Sciences and Arts Northwestern Switzerland  
Olten, Switzerland  
and  
Center for Adaptive Security Research and Applications (CASRA)  
Zurich, Switzerland

Adrian Schwaninger  
School of Applied Psychology  
University of Applied Sciences and Arts Northwestern Switzerland  
Olten, Switzerland  
and  
Center for Adaptive Security Research and Applications (CASRA)  
Zurich, Switzerland

Abstract—Initial and recurrent certification of aviation security officers (screeners) is mandatory in Europe as specified in the EU regulation 185/2010 and in national regulations. The certification includes a practical test, an X-ray image interpretation test as well as a theoretical test. This paper focuses on theoretical tests which are used to evaluate whether screeners have the relevant knowledge about regulations, security control procedures, etc. To test this knowledge multiple choice (MC) tests can be an efficient, reliable, valid and standardized method. The development of a MC test does not only require expertise in aviation security but also expertise in test development and psychometrics. Besides the standardized development of MC questions, the scoring method is likewise important. For this pilot study a first version of 188 MC questions from several EU Member States were evaluated with 183 screeners. The results revealed that it is of utmost importance that the development of MC questions is based on MC guidelines, i.e. apply expertise in test development and psychometrics in order to guarantee a high quality. Besides the test development the scoring method has a substantial influence on test difficulty, reliability, item difficulty and item discrimination. Implications of the three different scoring methods were discussed.

Keywords—airport security, certification tests, theoretical tests, multiple choice tests

I. INTRODUCTION

The terrorist attacks of September 11, 2001 changed the world of civil aviation. Since then several successful and attempted terrorist attacks were conducted and airports are confronted with new threat dimensions. In order to cope with this new situation, modern technologies were developed and implemented. Despite of the increasing automation of security equipment, the final decision is still made by the human operator. The best technology is of limited value if the airport security officers (screeners) who operate it are not selected and trained appropriately. In order to evaluate the relevant competencies of screeners, initial and recurrent certification tests are mandatory according to the Commission Regulation of the European Union (EU) No 185/2010 and national regulations. Certification is a formal evaluation and confirmation by or on behalf of the appropriate authority indicating that the person has successfully completed the relevant training and that the person possesses the necessary competencies to perform assigned functions to an acceptable level. The certification of a screener includes the following test types: practical tests, X-ray image interpretation tests and theoretical tests.

Practical tests are used to evaluate for example whether a screener knows how to operate the X-ray machine or how a proper pat down search of passengers has to be performed. X-ray
image interpretation tests measure whether a screener has the necessary X-ray image interpretation competency. Theoretical tests are used to evaluate whether airport security officers have the relevant knowledge about regulations, security control procedures, equipment and its use, etc. There are several possibilities to measure the theoretical knowledge of screeners, e.g. multiple choice (MC) tests, open-ended questions, gap tests. Very often MC tests are used to measure the theoretical knowledge as they are an efficient and objective way to assess knowledge [1].

A MC question always consists of a stem and a varying number of correct answers plus a varying number of wrong answers which are called distractors. The development of MC tests for certification of screeners does not only require subject matter expertise in aviation security but also expertise in test development and psychometrics. This is of particular importance, as passing the certification is a prerequisite for working or continuation of work as a screener and therefore the test has to fulfill test psychological criteria, such as for example reliability. Reliability indicates whether a test is consistent and the results are repeatable. Consistency means that the items of a test should all measure the same competency – in this case the theoretical knowledge of a screener. Repeatability means that the test should yield approximately the same result today and on a later test occasion (provided that the screener did not train between the tests). According to [2] there are different methods of estimating reliability, such as test-retest, alternate forms, split-half and internal consistency. In this study Cronbach’s alpha and the Kuder-Richardson equation were used to test the internal consistency. Further measures (difficulty measures, distractor measures and item discrimination measures) may be used to understand why a test shows specific levels of reliability [2].

Beside the standardized development of MC questions, the scoring method is likewise important. There are different methods to evaluate MC questions that vary in the way how credit is awarded to the test taker. Using the Multiple Response (MR) method, each question is treated as an entity and a point is awarded only if all the response options of the question were marked correctly, see e.g. [3]. The MR scoring method is based on the assumption that reduced probability of chance performance (i.e. performance that would be achieved by randomly answered questions) results in increased reliability of the test [4]. A characteristic feature of the MR method is that partial knowledge is ignored, i.e. if a person’s response is only partially correct, the entire response will be ignored regardless of its potential for containing useful information. There is evidence that taking into account partial knowledge when calculating test scores will lead to better psychometric properties of a test [5].

The other extreme is the Multiple True False (MTF) scoring method where credit is apportioned evenly for all true-false options answered correctly [3]. The basic principle underlying MTF is the assumption that each response of an examinee contains valid information, even for chance performance. MR and MTF are the methods that are located at the maximum respectively the minimum position on the continuum how strict credit is given.

Between these two extreme methods there are other methods which do not credit each correctly answered option but still award partial points. These methods differ in the minimum number of correctly answered options that lead to partial points. To cover this approach within our study we chose one of these scoring options: Method#2. This method is characterized by the fact that it only credits correct answers if the total number of correct answered options is above chance performance.

The scoring method seems to have a large influence on the test difficulty which in turn should be in line with the minimum score a screener has to achieve in order to pass the test (i.e. the definition of the pass mark).

The first aim of this study was to evaluate the quality of the existing item pool of EU member states. The second aim was to evaluate whether the scoring method influences not only the test difficulty, but also psychometric properties such as test reliability, item difficulty and item discrimination. Based on these findings, recommendations are provided for the development, evaluation and implementation of theoretical tests for the certification of screeners.

II. METHOD AND PROCEDURE

A. Participants

The 183 participants were all screeners of three different European airports. 46.1% of them were females and 53.9% were males.

B. Material

The questions used for this study were provided by several EU Member States. The primary item pool was created by aggregating questions covering the same topic which resulted in a final item pool including a total of 188 questions: 58 questions in the basic part and 130 questions in the Cabin Baggage Screening (CBS) part. The basic part included questions which are independent of the working area and had to be answered by all screeners, such as questions about access authorization, the meaning of colors in X-ray images, etc. The CBS part included questions which had to be answered by CBS screeners only, such as what kind of items are prohibited for passengers, questions about the walk through metal detector, etc.

C. Procedure

All screeners completed the basic part of the test and 178 of them also took the CBS part. As concentration can decrease over time, screeners who had to conduct both parts were asked to take the second part on the next day. Each participant received a token, which allowed conducting the test anonymously and which was used to give an individual feedback afterwards.

231
Each screener completed the MC test in the form of a simple online questionnaire accessible via web browser. During the test, participants were not able to review their answers and did not receive any feedback.

III. RESULTS AND DISCUSSION

As already mentioned in the introduction, for the test evaluation the MR scoring method, the Method#2 and the MTF scoring method were used.

With the MR method one item is treated as an entity and one point is awarded only if all the response options of an item were marked correctly [3]. The MR scoring method is based on the assumption that reduced probability of chance performance results in increased reliability of the test [4]. With this method, answering some response options correctly but not all, counts as if none were answered correctly. Correctness according to the MR scoring method is as follows:

$$Correctness = \frac{\sum_{i=1}^{N} \left( \begin{array}{l} 1, \text{if } \frac{c_i + d_i}{m_i} = 1 \\ 0, \text{if } \frac{c_i + d_i}{m_i} = 0 \\ \end{array} \right)}{N}$$

(1)

where  
- \(c_i\): Number of correct answers found in question \(i\)
- \(d_i\): Number of correct recognized distractors in question \(i\)
- \(m_i\): Number of choices in question \(i\)
- \(N\): Number of questions

The second scoring method, the Method#2 takes into account partial knowledge by awarding half credit for three options correct and full credit for all correct [5]. Correctness according to the Method#2 is as follows:

$$Correctness = \frac{\sum_{i=1}^{N} \left( \begin{array}{l} 1, \text{if } \frac{c_i + d_i}{m_i} = 1 \\ \frac{1}{2}, \text{if } m_i - 4 \text{ and } c_i + d_i = 3 \\ 0, \text{otherwise} \\ \end{array} \right)}{N}$$

(2)

where  
- \(c_i\): Number of correct answers found in question \(i\)
- \(d_i\): Number of correct recognized distractors in question \(i\)
- \(m_i\): Number of choices in question \(i\)
- \(N\): Number of questions

Another scoring method is the MTF scoring method which apportions credit evenly for all true-false options answered correctly [3]. The basic principle underlying MTF is the assumption that each response contains valid information, even for performance below chance performance. Correctness according to the MTF scoring method is as follows:

$$Correctness = \frac{\sum_{i=1}^{N} \frac{c_i + d_i}{m_i}}{N}$$

(3)

where  
- \(c_i\): Number of correct answers found in question \(i\)
- \(d_i\): Number of correct recognized distractors in question \(i\)
- \(m_i\): Number of choices in question \(i\)
- \(N\): Number of questions

A. Test Difficulty

Table I shows the average test difficulty and standard deviation of both test parts (basic and CBS) using the three different scoring methods.

<table>
<thead>
<tr>
<th>Test parts</th>
<th>Scoring Method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MR</td>
</tr>
<tr>
<td>Basic</td>
<td>M=0.53</td>
</tr>
<tr>
<td>CBS</td>
<td>M=0.64</td>
</tr>
</tbody>
</table>

Analyses of variance (ANOVA) using performance scores of the three scoring methods as within participant factor (MR, Method#2, MTF) were conducted for the basic and the CBS part, respectively. There were large main effects of scoring method for the basic \(F(2, 546) = 946, p < .001\), and the CBS part \(F(2, 531) = 595, p < .001\). Furthermore, the multiple comparison test by Tukey revealed significant differences between the MR method and the Method#2 \((p < .001)\), between the MTF and Method#2 \((p < .001)\) as well as between the MTF and the MR method \((p < .001)\).

These results were found for both parts, the basic and CBS part. As expected, the MR method with a high item difficulty results in lower overall performance compared to the MTF method and Method#2. Another difference between the three scoring methods is the variance which can be described by the standard deviation between screeners. According to Table I the MR method shows the highest variance. That is, the greatest differentiation between screeners was obtained with the MR scoring method. Furthermore, the performance scores obtained in the CBS part, which aimed at verifying the knowledge of those screeners with knowledge gaps, there should be items of varying difficulty for screening purposes and these should be excluded from the item pool if the advisable range of difficulties, which is 

$$\text{SCBS} = \frac{\sum_{i=1}^{N} \left( \begin{array}{l} 1, \text{if } \frac{c_i + d_i}{m_i} = 1 \\ 0, \text{otherwise} \\ \end{array} \right)}{N}$$

AVERAGE TEST DIFFICULTY AND STANDARD DEVIATION

<table>
<thead>
<tr>
<th>Test parts</th>
<th>Scoring Method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MR</td>
</tr>
<tr>
<td>Basic</td>
<td>M=0.53</td>
</tr>
<tr>
<td>CBS</td>
<td>M=0.64</td>
</tr>
</tbody>
</table>

As expected, the MR method with a high item difficulty results in lower overall performance compared to the MTF method and Method#2. Another difference between the three scoring methods is the variance which can be described by the standard deviation between screeners. According to Table I the MR method shows the highest variance. That is, the greatest differentiation between screeners was obtained with the MR scoring method. Furthermore, the performance scores obtained in the CBS part, which aimed at verifying the knowledge of those screeners with knowledge gaps, there should be items of varying difficulty for screening purposes and these should be excluded from the item pool if the advisable range of difficulties, which is 

$$\text{SCBS} = \frac{\sum_{i=1}^{N} \left( \begin{array}{l} 1, \text{if } \frac{c_i + d_i}{m_i} = 1 \\ 0, \text{otherwise} \\ \end{array} \right)}{N}$$

AVERAGE TEST DIFFICULTY AND STANDARD DEVIATION

<table>
<thead>
<tr>
<th>Test parts</th>
<th>Scoring Method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MR</td>
</tr>
<tr>
<td>Basic</td>
<td>M=0.53</td>
</tr>
<tr>
<td>CBS</td>
<td>M=0.64</td>
</tr>
</tbody>
</table>

As expected, the MR method with a high item difficulty results in lower overall performance compared to the MTF method and Method#2. Another difference between the three scoring methods is the variance which can be described by the standard deviation between screeners. According to Table I the MR method shows the highest variance. That is, the greatest differentiation between screeners was obtained with the MR scoring method. Furthermore, the performance scores obtained in the CBS part, which aimed at verifying the knowledge of those screeners with knowledge gaps, there should be items of varying difficulty for screening purposes and these should be excluded from the item pool if the advisable range of difficulties, which is 

$$\text{SCBS} = \frac{\sum_{i=1}^{N} \left( \begin{array}{l} 1, \text{if } \frac{c_i + d_i}{m_i} = 1 \\ 0, \text{otherwise} \\ \end{array} \right)}{N}$$

AVERAGE TEST DIFFICULTY AND STANDARD DEVIATION

<table>
<thead>
<tr>
<th>Test parts</th>
<th>Scoring Method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MR</td>
</tr>
<tr>
<td>Basic</td>
<td>M=0.53</td>
</tr>
<tr>
<td>CBS</td>
<td>M=0.64</td>
</tr>
</tbody>
</table>

As expected, the MR method with a high item difficulty results in lower overall performance compared to the MTF method and Method#2. Another difference between the three scoring methods is the variance which can be described by the standard deviation between screeners. According to Table I the MR method shows the highest variance. That is, the greatest differentiation between screeners was obtained with the MR scoring method. Furthermore, the performance scores obtained in the CBS part, which aimed at verifying the knowledge of those screeners with knowledge gaps, there should be items of varying difficulty for screening purposes and these should be excluded from the item pool if the advisable range of difficulties, which is 

$$\text{SCBS} = \frac{\sum_{i=1}^{N} \left( \begin{array}{l} 1, \text{if } \frac{c_i + d_i}{m_i} = 1 \\ 0, \text{otherwise} \\ \end{array} \right)}{N}$$

AVERAGE TEST DIFFICULTY AND STANDARD DEVIATION

<table>
<thead>
<tr>
<th>Test parts</th>
<th>Scoring Method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MR</td>
</tr>
<tr>
<td>Basic</td>
<td>M=0.53</td>
</tr>
<tr>
<td>CBS</td>
<td>M=0.64</td>
</tr>
</tbody>
</table>
screeners who specifically screen passengers, baggage and items carried were better than those obtained in the basic part. As this effect is only existent for the MR method the participants might have felt more comfortable with questions closely related to their job, rather than with general questions about aviation security.

B. Reliability

To examine whether the two test parts which were used to measure the theoretical knowledge are reliable and may be used for certification purposes an item analysis was conducted. The internal consistency of the test was estimated using Cronbach’s alpha respectively the Kuder-Richardson equation for the MR scoring method (since this method results in dichotomous data). The following Table II gives an overview of the results. As a commonly used rule of thumb regarding reliability, correlation coefficients of at least 0.75 and Cronbach’s alpha of at least 0.85 are recommended (for further information see [6]).

<table>
<thead>
<tr>
<th>Test parts</th>
<th>Scoring Method</th>
<th>MR</th>
<th>Method#2</th>
<th>MTF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic</td>
<td>MR</td>
<td>0.68</td>
<td>0.71</td>
<td>0.72</td>
</tr>
<tr>
<td>CBS</td>
<td>MR</td>
<td>0.81</td>
<td>0.78</td>
<td>0.78</td>
</tr>
</tbody>
</table>

As can be seen in Table II the used version of both test parts did not achieve the required minimum standard of test reliability. This is not unusual, since in most cases MC questionnaires have to be pilot-tested and refined several times to achieve good reliability scores. Furthermore, not all MC questions in the test were developed using MC guidelines to guarantee a high quality of MC questions and hence a sufficient reliability score.

Subsequently, item difficulty and item discrimination were analyzed in order to get further information on each MC question and an indication why the reliability scores were on a rather low level.

C. Item Difficulty

The item difficulty is a parameter that shows how many participants answered the question correctly. An item is very difficult if nearly nobody was able to solve it, whereas it is very easy if most of the test takers marked the correct answers. To discriminate between screeners with very good knowledge and screeners with knowledge gaps, there should be items of varying difficulties with values between 0.2 and 0.8. For more information regarding item difficulty see [2] or [7]. Fig. 1 shows the distribution of item difficulties for all three scoring methods and both test parts.

Item difficulty for both test parts (basic and CBS) when calculated on the basis of the MTF method, show an unbalanced distribution of the values. Most questions have values of 0.8 or even higher. As one would expect, the MC questions are quite easy if partially correct answers are taken into account. If the item difficulties are calculated based on scoring Method#2 the distribution is higher, but item difficulties are still disproportionately high. Only the MR scoring method shows a distribution over the whole range of values and most of the item difficulties are covered.

Table III presents the number of items that should be excluded from the item pool if the advisable range of difficulties, i.e. between 0.2 and 0.8 is to be realized. The number of items out of the advisable range is much higher with the MTF method and Method#2 compared to the MR method. This trend applies to the basic part as well as to the CBS part.

<table>
<thead>
<tr>
<th>Test parts</th>
<th>Difficulty</th>
<th>Scoring Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic</td>
<td>MR</td>
<td>Method#2</td>
</tr>
<tr>
<td>&lt; 0.2</td>
<td>14</td>
<td>4</td>
</tr>
<tr>
<td>&gt; 0.8</td>
<td>12</td>
<td>31</td>
</tr>
<tr>
<td>Total</td>
<td>26</td>
<td>35</td>
</tr>
<tr>
<td>CBS</td>
<td>MR</td>
<td>Method#2</td>
</tr>
<tr>
<td>&lt; 0.2</td>
<td>52</td>
<td>2</td>
</tr>
<tr>
<td>&gt; 0.8</td>
<td>10</td>
<td>66</td>
</tr>
<tr>
<td>Total**</td>
<td>62</td>
<td>68</td>
</tr>
</tbody>
</table>

---

*total of excluded items out of an item pool of 58
**total of excluded items out of an item pool of 130
D. Item Discrimination

Furthermore, the item discrimination was evaluated. The item discrimination examines how precisely each item reflects the overall result of the test takers [2]. Consequently, an item with a high discriminating power shows a high correlation with the overall result so that a test taker with a very good overall result also has a high score in the single item. On the contrary, items with a low or even a negative item discrimination value do not measure the same competence as the test aims to.

According to [7] the advisable ranges are item discrimination values between 0.3 and 1. In an area with a very wide and differentiating subject, the limit of 0.3 is very high, hard to reach and therefore a lower limit is acceptable and often used.

Even if the difference between the item discrimination values are only small for the three scoring methods, it has an effect on the number of questions that are below the advisable range and therefore influences the number of items that are useful for the test. Table IV shows how many items need to be excluded from the test if the item discrimination limit is lowered to 0.15 for the basic and CBS part. It may be assumed that the item discrimination values would have been higher if all MC questions had been developed according to the existing general guidelines [8].

TABLE IV. NUMBER OF ITEMS THAT SHOULD BE EXCLUDED BASED ON ITEM DISCRIMINATION

<table>
<thead>
<tr>
<th>Test parts</th>
<th>Item discrimination</th>
<th>Scoring Method</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic</td>
<td>&lt; 0.15</td>
<td>MR 20*</td>
<td>Method#2 23*</td>
</tr>
<tr>
<td>CBS</td>
<td>&lt; 0.15</td>
<td>MTF 20*</td>
<td>MTF 65**</td>
</tr>
</tbody>
</table>

**total of excluded items out of an item pool of 130
*total of excluded items out of an item pool of 58

IV. Conclusions

This study compared three MC scoring methods empirically regarding their impact on test difficulty and their ability to differentiate screeners’ aviation security knowledge. Average performance is much lower with the MR method compared to Method#2 and the MTF method. Further, the standard deviation is significantly higher with the MR method compared to MTF and Method#2 which results in a much better differentiation between screeners with good knowledge and those with knowledge gaps, why the MR method may be favored.

Reliability values were all below the recommended minimum of 0.85 irrespective of the scoring method that was applied. The MR method showed reliability values which were both the highest for the CBS part as well as the lowest for the basic part. In the CBS part the reliability value supports rather the assumption from [4] that reduced probability results in increased reliability whereas the reliability value in the basic part supports findings from [5] or [9]. Overall, the results clearly showed that the item pool needs to be further improved taking into account elaborated guidelines to develop MC questions as for example provided by [8], [10] or [11] in order to reach acceptable reliability values and to be able to reasonably evaluate the different scoring methods.

As expected, the different scoring methods had also effects on the item parameters such as the item difficulty and the item discrimination. When using the MR method, item difficulty varied between very easy and very difficult. With the other two methods most of the questions resulted in very low difficulty scores. Yet, even if the MR method were to be used, half of the items would have to be revised because they are either too easy or too difficult. Concerning the item discrimination, the values show rather small differences, which nevertheless have an effect on the number of items that should be excluded from the item pool. Using the MR method 72 questions out of 130 should be revised compared to 80 questions respectively 85 questions if Method#2 or the MTF method is applied.

To summarize, this pilot test showed that many questions of the existing item pool need to be adapted as the item difficulty and/or the item discrimination do not comply with the advisable ranges. Dependent on the scoring method, between 61% and 82% of the questions do not correspond with the recommended range of the relevant statistical values. Such an effect occurs especially if MC questions are developed without the help of guidelines respectively expert knowhow in the development of MC questions. Furthermore, MC tests often have to be revised at least once, in the majority of cases even a number of times in order to achieve sufficient reliability values. Revisions should be done based on the values of the item difficulty and the item discrimination: Items that are either too easy or too difficult do not differentiate between persons with good knowledge and persons with knowledge gaps. Likewise items with a low item discrimination value are of no use for the test, as they do not measure the intended construct that the test was designed to measure.

Furthermore, the results in this study showed that the MR method may possibly be better if appropriate item difficulty and item discrimination parameters have to be achieved which in turns might have a positive effect on the reliability score. However, as already mentioned further studies should be done.

Beside these aspects arising from the statistical analysis also practical consequences have to be considered. Using the MR method implies that every single option in a MC question and thus each step in the process must be correct. If one detail is wrong the whole reaction might be wrong which means that partial points are not acceptable. The MR method automatically results in a higher test difficulty and thus a lower pass mark has to be defined. In case that partial knowledge is to be considered and thus the MTF method is preferred, higher average scores can be expected and thus, pass marks can be higher. Moreover, using the MTF method more questions might be required to replace...
questions with insufficient item difficulty and item discrimination values in order to achieve an acceptable reliability.

This pilot study revealed that there are several necessary steps to be taken in order to obtain a reliable instrument for the certification of screeners using a theoretical test. Firstly, the development of MC questions needs to be based on guidelines [8]. Secondly, the appropriate scoring method has to be chosen and the pass mark must be defined in accordance with the scoring method. Thirdly, the reliability of a test has to be evaluated. However, other test criteria such as objectivity, validity and standardization must not be neglected. In order to evaluate these criteria it is of utmost importance to conduct a pilot study. Based on this pilot study a first revision of the test should be conducted. In most cases several revisions are necessary until the test can be used as a certification test.

ACKNOWLEDGMENT

We would like to thank the European Civil Aviation Conference (ECAC) for supporting this study. We further acknowledge the support rendered by the European Airports providing us with highly qualified personnel in the area of aviation security.

REFERENCES


