International Journal of Assessment Tools in Education: Vol. 3, Issue 2 (2016) pp. 137-150

http://www.ijate.net
e-ISSN: 2148-7456 © IJATE
INTERNATIONAL JOURNAL of ASSESSMENT TOOLS in EDUCATON

"Original Research Article"

Is This Reliable Enough? Examining Classification Consistency and Accuracy in a Criterion-Referenced Test

Susanne Alger[1,*]

[1]Department of Applied Educational Science, Umeå University, SE-901 87 UMEÅ, Sweden

Abstract

One important step for assessing the quality of a test is to examine the reliability of test score interpretation. Which aspect of reliability is the most relevant depends on what type of test it is and how the scores are to be used. For criterion-referenced tests, and in particular certification tests, where students are classified into performance categories, primary focus need not be on the size of error but on the impact of this error on classification. This impact can be described in terms of classification consistency and classification accuracy. In this article selected methods from classical test theory for estimating classification consistency and classification accuracy were applied to the theory part of the Swedish driving licence test, a high-stakes criterion-referenced test which is rarely studied in terms of reliability of classification. The results for this particular test indicated a level of classification consistency that falls slightly short of the recommended level which is why lengthening the test should be considered. More evidence should also be gathered as to whether the placement of the cut-off score is appropriate since this has implications for the validity of classifications.

Article Info

Received
15 January 2016

Revised:
23 March 2016

Accepted
10 April 2016

Keywords:
reliability, criterion-referenced test, driving licence test, classification consistency, decision consistency, single administration

1. INTRODUCTION

Test scores are often used as the basis for decisions of various kinds. To determine to what extent such a use of the score is appropriate the validity and reliability of the score interpretation must be assessed. Validity issues concerns to what degree evidence can be found to support that test scores actually reflect the construct intended to be measured and whether the proposed use of the score is appropriate. To make defendable interpretations of test scores possible the scores need to be reliable. If the test result reflects random error to a high degree rather than the construct intended to be measured the potential for wise decisions is limited (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014).

Reliability concerns the precision of test scores, but, like validity, must be viewed in the context of intended test use and interpretation. Not all reliability measures can be used both for scores used to compare the test-takers to each other or to a norm group (norm-referenced)

[*] Corresponding Author Phone: +46 907866221 Fax: +46 907866686
E-mail: susanne.alger@umu.se

2148-7456 /© 2016 DOI: 10.21449/ijate.245198

and scores used to determine the achievement level of the test-taker in relation to a particular criterion (criterion-referenced). For tests where the end result is a classification, conventional measures of reliability may not be suitable (Brennan, 2006). In the case of a criterion-referenced certification test the consistency of interest often concerns the classification (e.g. pass or fail) rather than the individual score. Reliability of classification can be described in terms of classification consistency or classification accuracy. *Classification consistency* concerns to what extent test-takers are consistently reported as having passed or failed. An agreement coefficient represents the proportion of consistently classified tests. *Classification accuracy* is often expressed in terms of false positive and false negative error rates and concerns to what extent the classification reflect the test-taker's true score (Lee, 2010). Various methods can be used to calculate classification accuracy (see e.g. Rudner, 2005 and Guo, 2006). The true score is a theoretical construct representing the mean of an unlimited number of test performances by the same individual. Since we cannot know what the actual true score really is an estimation is made based on observed test data (Crocker & Algina, 1986).

The concept of reliability is connected to replications in a broad sense, be it over time, test versions or items, where certain aspects are, or are intended to be, the same. The replication does not have to be an actual full replication, but in the case of a hypothetical replication certain assumptions have to be made (Brennan, 2006). Although it is recommended in the Standards for Educational and Psychological Testing (AERA, APA, & NCME, 2014) that two administrations are used and many reliability measures are based on two test administrations (see e.g. Berk, 1980; Hambleton, Swaminathan, Algina, & Coulson, 1978) that is not always feasible and the single-administration situation is the norm rather than the exception. For tests with only one administration it may be less straightforward to estimate the classification reliability but this does not make the issue less important. There are a number of single-administration procedures for estimating classification consistency. Some are limited to tests consisting of equally weighted, dichotomously scored items (e.g. Hanson & Brennan, 1990; Huynh, 1976; Peng & Subkoviak, 1980; Subkoviak, 1976, 1988). To deal with polytomous items, mixed formats and composite scores other procedures have been developed. (e.g. Livingston & Lewis, 1995; Woodruff and Sawyer,1989; Breyer & Lewis, 1994; Brennan and Wan, 2009). If we broaden the scope to include IRT models the number increases further (e.g. Han & Hambleton, 2007; Huynh, 1990; Lathrop, 2015; Lee, 2010; Lee, Hanson, & Brennan, 2002; Wang, Kohlen & Harris, 2000; Wyse & Hao, 2012).

When evaluating tests the concept of reliability should not be reduced to, or confused with, a particular method. There are advantages and drawbacks to all methods for estimating classification consistency and accuracy, which is why it would be good to use several. When choosing a reliability measure the type of test and the use of its result must be taken into consideration. In this case it was decided that classical test theory was a suitable first approach as it is useful for straightforward multiple choice tests (Schuwirth & van der Vleuten, 2011). Also, as the administration and scoring of the test examined here is standardized, classical test theory was chosen over e.g. generalizability theory (where several sources of error is evaluated through extensive computation). IRT modelling would have been possible as there are many respondents, but that requirement may limit its usefulness for some practitioners and as one of the ambitions with the present study was to apply analyses that would be reasonably easy to use for practitioners and stake holders working with similar data, classical test theory methods were applied in this particular study.

Even within classical test theory various approaches have been used for assessing single-administration classification consistency. Livingston and Lewis (1995) replaced the original test with an idealized test with dichotomously scored items. Woodruff and Sawyer

(1989) and Breyer and Lewis (1994) developed split-half methods whereas Brennan and Wan (2009) developed a bootstrap procedure. Huynh (1976), Subkoviak (1976, 1988), and Hanson and Brennan (1990) have all developed methods for dichotomously scored equally weighted items and in the present study, Subkoviak's and Hanson and Brennan's methods were chosen, due to the availability of software and tables (and the possibility to replicate a previous study in the same test context that used Subkoviak's index, see Sundström, 2003; Wiberg, 2004). These methods are therefore used here as a first step to examine three versions of a criterion-referenced licensing test – the theory part of the Swedish driving licence test. This test is administered to thousands of people every week and there are a number of alternate forms.

The driving test is a high-stakes test both from the perspective of the individual and of society as a whole. Nevertheless very little has been done to examine this test in terms of the reliability of the classification. This situation does not seem to be unique for Sweden. Considering that all over the world there are millions of people tested in this context in some manner, be it a test for permission to start driving practice or some oral questions after passing the practical driving test, it is surprisingly difficult to find studies of the quality of driving licence testing (see however Baughan & Simpson, 1999; Henriksson, Sundström, & Wiberg, 2004; Reiner & Hagge, 2006; Siegrist, 1999; Wiberg & Sundström, 2009).

A previous study of inter-rater reliability of the Swedish practical driving test showed a 93 per cent agreement rate for pass/fail distinction when 83 examiners were accompanied by one out of five supervising examiners for a day, covering 535 tests in total (Alger & Sundström, 2013). Apart from traditional reliability indices such as measures of internal consistency, no reliability studies have been carried out on the current test, i.e. since the test last underwent a significant revision in 2006 (although Sundström, 2003, and Wiberg, 2004, examined classification consistency of a previous version of the theory test). As the quality of driving licence tests is as deserving of attention as the quality of many other achievement tests, and warrants the same evaluation, this lack of published work makes the national example presented in this paper all the more important.

The purpose of this study is to examine three versions of the theory part of the Swedish driving licence test to estimate to what extent they can be regarded as consistent in classification (in terms of pass/fail) and accurate in terms of classification.

2. METHOD

2.1. Participants

All data were collected within a certain time period, but not from the same test occasion. First-time test takers were selected in order to avoid having several tests by the same person influence results. The sample included 12,072 test-takers who had taken one of three versions of the test. Table 1 shows that there were slightly more women than men and the majority had registered for the test via a driving school. The age distribution is positively skewed (between 2.02 and 3.28 for the three test versions). The minimum age for taking the driving licence test is 18, but ages in the sample range to 72. However, almost half of the test-takers were 18-year-olds.

Table 1. Descriptive statistics for background variables for first time test-takers

Test version	*N*	Women (%)	Driving school (%)	Age		
				M	*SD*	*MD*
1	3,947	51.6	60.3	21.2	6.2	18
2	4,017	52.6	58.4	21.2	6.0	19
3	4,108	52.2	59.0	21.0	5.9	18

There are no statistically significant differences between the test-takers taking the three selected test versions when it comes to age $F(2, 12{,}069) = 1.91$, $p = .148$, , gender, $\chi^2(2, N=12{,}072) = .858$, $p = .651$, and method of registration for the test (via a driving school or not), $\chi^2(2, N=12{,}072) = 3.005$, $p = .223$, – variables known to affect results – so even though not the same group has taken all three forms, or a representative selection of items in all forms, any differences in results are likely to be due to the test versions.

2.2. Instruments and Procedure

The Swedish driving licence test consists of two parts – theory and practical driving. Both test parts are booked at the same time and carried out on the same day or within a few days of each other. The test-taker has to pass both tests within a two-month period to obtain a driving licence. There are no specified limits as to how many times the test can be retaken.

The theory test is computerized and distributed by the Swedish Transport Administration in many different locations. It consists of 65 multiple choice items (plus five try-out items). The number of response options varies from two to six, four being the most common. The order of the items and the order of the options are randomized to make cheating harder.

The try-out questions are not awarded points, but out of the remaining 65 items the test-taker has to answer 52 items correctly (80%) within a 50 minute period in order to pass. The result is given on the screen once the test-taker has completed the test. There is no scaling so the raw score is the final score. The conditions under which the theory tests are administered are standardized. The process for test scoring is automatic and identical for all tests. These efforts to avoid random error can be viewed both as promoting fairness and increasing test reliability.

248,546 theory tests were administered in 2012 (86 per cent of them in Swedish). There were 66 different test versions (some have items in common). For this study three test versions from 2012 were selected which do not have any items in common. The test versions selected for this study were all administered between the 3rd of July and the 17th of August and were the test versions distributed to the largest number of test-takers in 2012. Tests that were distributed in other languages or distributed in Swedish but translated orally by an interpreter were not included in the following analyses.

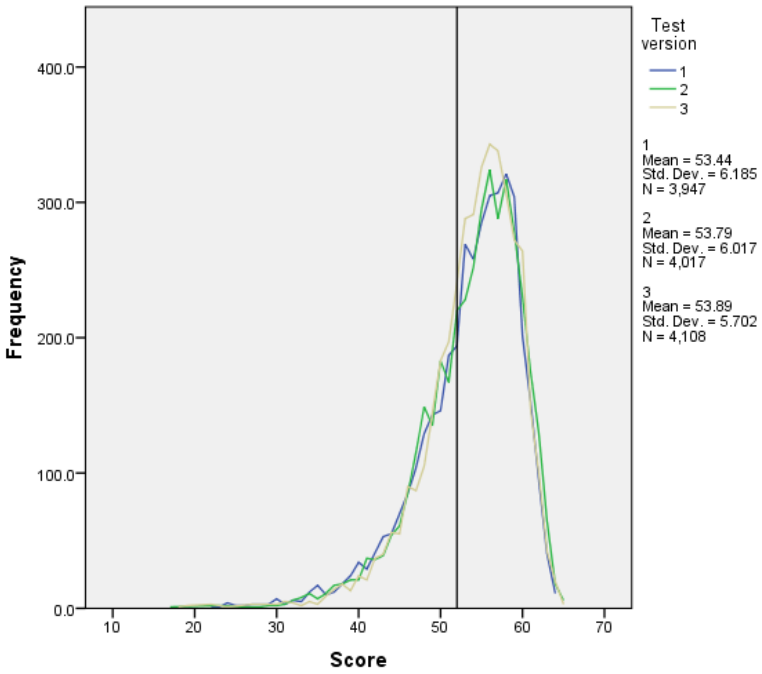


Figure 1. Score distribution for the three test versions in relation to the cut-off score of 52.

The medians are the same for all three test versions (55 points). The mean scores and variance of the test versions are not identical, but similar. Version 1 has a lower mean score than the other two versions $F(2, 12{,}069) = 6.35$, $p = .002$, $\eta^2 = 0.001$. That the difference is statistically significant is mainly due to the size of the sample rather than the size of the difference as the effect size is very small. Even though the differences between means are small the mean is so close to the cut-off score that the percentage of students who pass varies significantly between versions $\chi^2(2, N=12{,}072) = 10.055$, $p = .007$. The largest difference was between version 1 and 3, but all differences are small. The percentage of students who passed was 69.4, 70.3 and 72.5 for the three versions. Version 3 was the version where the highest percentage of fist time test-takers passed, i.e. scored 52 points or more. The percentage of correct responses for each item indicates that the level of difficulty (p-value) varies between items, but in a similar manner for all three versions (see Figure 2). Item results have been sorted by size and do not reflect the actual order of items.

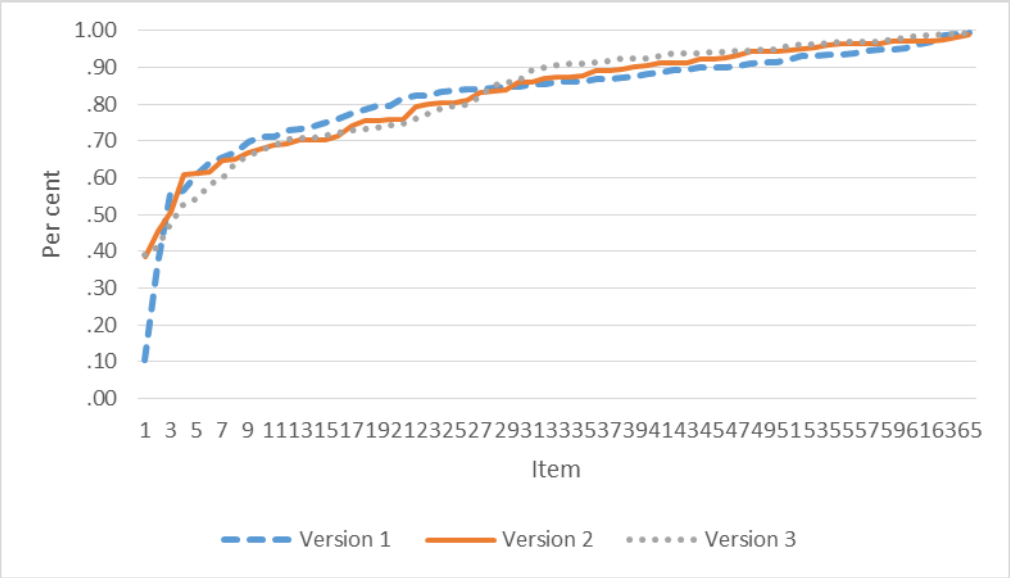


Figure 2. P-value for each item for the three test versions.

One item in version 1 proved considerably more difficult for the test-takers, but otherwise the three versions have fairly similar item results on average.

2.3. Data Analysis

As mentioned earlier hypothetical replications are based on certain assumptions. Subkoviak's model is based on the assumption that scores are independently distributed and his procedure calculates the index for one test-taker at a time and then averages the outcome. The Hanson-Brennan approach uses an estimated true score distribution. Subkoviak makes no assumptions for the observed score distribution whereas Hanson & Brennan assume a beta-binomial model for observed scores.

Assuming that the experience of taking the test X with cut-off score C would not affect the outcome of taking a parallel test (i.e. independent scores), that both score distributions for the individual are identically binomial in form and that all items have the same difficulty the coefficient of agreement for an individual test-taker with Subkoviak's method (Subkoviak, 1976) can be estimated by using Equation 1.

$$P_C^{(i)} = \left[P(X_i \geq C)\right]^2 + \left[1 - P(X_i \geq C)\right]^2 \quad (1)$$

where

$$P\left(X_i \geq C\right) = \sum_{X_i=C}^{n} \binom{n}{X_i} p_i^{X_i} \left(1 - p_i\right)^{n-X_i} \quad (2)$$

In Equation 2 p_i is an estimation of the true proportion-correct score. The coefficient of agreement P_C for a group of N persons is the mean of the individual coefficients obtained from Equation 1 (Subkoviak, 1976).

As Subkoviak in a later article (1988) provided lookup tables where the agreement coefficients could be determined with the help of a reliability coefficient and z-score our calculations have been compared with values determined from those lookup tables to see if the tables provide a suitable shortcut. Classification consistency and accuracy has been calculated in terms of the 4-parameter Hanson and Brennan index (Brennan, 2004). These single administration methods were selected since they are suitable for dichotomously scored items and were deemed to be accessible.

The Hanson and Brennan index has been calculated with the freely available software BB-Class (http://www.education.uiowa.edu/centers/casma/computer-programs). The formulas used are provided in the manual for BB-Class. Other calculations have been made with Excel and SPSS.

The Hanson & Brennan (1990) four-parameter model which is used to estimate the true-score distribution has two shape parameters and a lower and upper limit of the beta true score distribution. The conditional error distribution is here assumed to be binomial.

Hanson and Brennan (1990) found that results did not differ much when they used the binomial or the two-term approximation to the compound binomial error distribution. The beta-binomial model was originally intended for test-items of equal difficulty, but even when this is not the case the models seem quite robust (Huynh & Saunders, 1980; Subkoviak, 1978).

3. RESULTS

The classification consistency and accuracy was examined for the three test versions using methods developed by Subkoviak and Hanson-Brennan. As mentioned previously, Subkoviak not only provides a formula for calculations but also lookup tables for estimating classification consistency in terms of an agreement coefficient. In table 2 coefficients obtained from the formula as well as from the lookup tables are presented.

When using Subkoviak's lookup tables (see Subkoviak, 1988) to estimate classification consistency, the value of KR20 was rounded to 0.8 for all test versions (from .80, .79 and .78). The Subkoviak calculation and the lookup tables give similar estimates for all three test versions.

Table 2. Agreement coefficients indicating probability of consistent classification

Test version	Agreement (p_0) from Subkoviak's tables[a]	Subkoviak calculation[b]	Hanson-Brennan 4-parameter
1	.80	.80	0.82
2	.80	.80	0.82
3	.81	.79	0.82

[a] Z-scores for the three version are 0.31, 0.38 and 0.42 respectively, which has been rounded to 0.3, 0.4 and 0.4 when interpreting Subkoviak's lookup table. [b]Subkoviak has outlined a few different estimators of the probability of a correct item response (Subkoviak, 1976). In the Subkoviak calculation in table 3 KR20 has been used when calculating the probability for a correct response for each person. If KR21 is used the agreement coefficient estimates are .01 lower and if the maximum likelihood estimator is used it adds .02 to the result.

In addition to Subkoviak's agreement coefficients the Hanson-Brennan coefficient was presented in Table 2. As shown in table 2 it is identical for the three versions, but at a slightly higher level than the Subkoviak coefficient. When the relationship between expected and actual observed scores was displayed in a graph (see Appendix 1) the fit of the 4 parameter model was deemed adequate.

The scores on the theory test together with the result on the driving test are used to make a decision, in this case whether a driving licence should be awarded (given that all other formal requirements are met). However there is always a certain risk that this is not an

accurate decision for the individual, and for binary decisions this uncertainty can be stated as probability of false-positive or false-negative error rates as in Table 3.

Table 3. Classification accuracy in terms of probability for correct and incorrect classification

Test version	Probability correct classification	False-positive rate	False-negative rate
1	.87	.06	.06
2	.87	.07	.06
3	.87	.06	.07

As shown in Table 3 the probability for a correct classification is the same for all three test versions (87 per cent). The probability for a test-taker with a true score below the cut-off score passing (false-positive) and a test-taker with a true score above the cut-off score failing (false-negative) is between six and seven per cent, regardless of version and type of error.

4. DISCUSSION

Classification consistency and classification accuracy are important aspects when interpreting scores from criterion-referenced tests, but are not always examined and reported. For driving licence tests, the applied example in this study, one would expect to find more studies focusing on these aspects, due to the consequences an erroneous classification can have in this context. Whether there is an actual lack of studies or just a lack of published work is unknown. The need to estimate classification consistency and classification accuracy is, however, far from unique for these tests. When it comes to criterion referenced tests, in particular certification tests, the classification issue should be considered as it is essential information for tests where the outcome is dependent on a cut-off score. Making a classification based on certain criteria is a complex process - criteria have to be selected and formulated, tests that try the right skills/performances/traits have to be designed and used in the appropriate manner. To make sure that the result is not a random number or letter combination on a piece of paper but actually reflects a meaningful distinction between those who demonstrate the desired competencies and those who do not, there has to be evidence for the reliability and validity of the interpretation of the score.

The purpose of this study was to examine three versions of the theory part of the Swedish driving licence test to estimate to what extent they can be regarded as consistent in classification (in terms of pass/fail) and accurate in terms of classification by using methods developed by Subkoviak (Subkoviak, 1976, 1988) and Hanson & Brennan (Brennan, 2004).

For the particular test studied here the measures for classification consistency examined, i.e. agreement coefficients, are similar for all three test versions. Around 80 per cent of test-takers are estimated to be consistently classified, which is comparable to results from studies of previous versions of the theory test (Sundström, 2003; Wiberg, 2004). Both procedures used to examine classification consistency - Subkoviak and Hanson-Brennan - give similar results, and so do Subkoviak's lookup tables. This would indicate that the lookup tables are a viable method for practitioners who want a quick indication of classification consistency.

So how should this information be interpreted? When discussing what value of the agreement coefficient is satisfactory Subkoviak mentions three aspects: test length, the importance of the decision and the proportion of masters and non-masters (i.e. test-takers who pass or fail). Tests used to make important decisions "should be sufficiently long to guarantee an agreement coefficient exceeding .85" (Subkoviak, 1988, p. 52). If less than ten per cent of

the test-takers fail the test, values around .95 can be expected. For a typical routine classroom assessment the test should reach an agreement coefficient of at least 0.75 if half the class is expected to fail and .85 if less than 15 per cent of test-takers fail the test. Thus, based on these guidelines it would be desirable to have an agreement coefficient higher than .8 in this case as the theory test examined here is a fairly long high-stakes test, which currently less than half of the test-takers fail.

Not only do we want alternate versions of the same test to give the same outcome over occasions (provided that the construct we are measuring has not changed) but also that the outcome is correct. The purpose of accuracy indices is to estimate how well the actual classification reflects true classification. As inconsistently classified results are bound to be correct some of the time the probability for a correct classification is often higher than the probability for a consistent classification, which is reflected in the assessment of classification accuracy in this case. Like many other certification tests, the test examined here is used to assure that a certain standard is attained. In this context one purpose of the test is to protect the public from drivers that are not competent. False positives, i.e. test-takers who pass the test despite not having reached the stipulated level of skills and knowledge, are therefore what is more worrying from a societal view. According to our analyses of classification accuracy the likelihood of a false positive is around six or seven per cent. If such a percentage were to be falsely passed as a result of these tests that would mean around 250 test-takers for each test version.

The critical issue when examining the reliability of a particular test is usually not whether one test is more reliable than another but what constitutes a defensible level of misclassification. This is very much dependent on context and consequences. False positives can only be entirely avoided if all test-takers are failed, which is an untenable solution, but measures can be taken to improve the situation. If the results are not sufficiently reliable for the intended purposes – are there other ways of obtaining such results? In our example passing the theory part is not the only requirement for obtaining a driving licence. As one of the reasons for having two distinct parts of the driving test is that they measure partly different qualifications it is presumptuous to assume that shortcomings of one part can be completely compensated by the other. Nevertheless, the fact that there is also the practical driving test reduces the likelihood of a false positive when it comes to granting an actual driving licence.

Although it is unrealistic to expect error free measurement it is worth considering how reliability can be improved. A more precise measurement would make erroneous classifications due to the error of the measurement less likely. In our example many test-takers want to try to pass the test as soon as it is even remotely possible. As a result the mean is often close to the cut-off score which makes misclassification more likely. The agreement index is larger when the mean is far from the cut-off score (Berk, 1980; Meyer, 2010), which is one reason why standard setting and the position of the cut-off score is a critical issue for criterion referenced tests.

Indices for classification consistency and accuracy depend on a well-placed cut-off score not only in terms of statistics. If the cut-off score does not reflect a suitable boundary between those who have the necessary qualifications and not then both the result and the consistency and accuracy of that decision loses meaning. This is not only a reliability issue, but very much a question of validity. If the claims only referred to a range of scores on the test then reliability evidence would be enough, but in practice there is always a stronger claim, referring to non-test situations. Since the cut-off score is the operationalization of the

performance standard, evidence for the appropriateness of this cut-off score is a vital component when assessing the quality of the test. A more in-depth discussion of requirements and standard setting is, however, beyond the scope of this article.

Test reliability can also be improved by higher item quality since badly constructed test items are a source of error in all kinds of tests. Item writers should be well trained and adhere to researched and established principles for item writing (see e.g. Haladyna, Downing, & Rodriguez, 2002). An ideal certification test would consist of items which all differentiate between those who have the required knowledge and not (while at the same time covering all the necessary content and fulfilling all other requirements for good test items). Unless the construct measured is very well defined and rather narrow so that such a boundary is obvious this is a very difficult goal to achieve, but certainly worth striving for.

The reliability of classification can not only be improved through higher item quality, but also through increased item quantity. Lengthening the theory test would improve reliability as it is presented here (provided that the quality of items remained the same and no new aspects of error were introduced), but there is a limit where the gains are not of such a magnitude that further increases are worth making. Increasing test length would mean increasing test time and having to produce more items. In the case of the theory test perhaps some sort of adaptive testing could be used to avoid increasing general test time (Wainer, 2000; van der Linden & Glas, 2010).

This study showed that the theory test did not quite reach an acceptable level of classification consistency and classification accuracy. Improving item quality as well as lengthening the test should seriously be considered in order to increase reliability. The score distribution also affects the measures for classification consistency, but, for other than statistical reasons, it would be better if test-takers in general performed considerably better than the cut-off score than even worse.

The analyses made here are a first step towards identifying classification issues and improving test quality. As the choice of reliability measure will impact both the outcome of the analysis and the interpretation of test results future studies of classification consistency should include IRT and perhaps generalizability studies based on these tests. Similarity between test versions in terms of item content and the rationale behind the cut-off score also need to be examined further. Naturally there are many other aspects of test quality to study and improve too, but when it comes to differentiating between those who have the necessary skills to safely handle a vehicle in traffic and those who have not is important to make sure that the tests for this purpose are reliable. This study indicates that there is room for improvement in this respect.

5. REFERENCES

Alger, S., & Sundström, A. (2013). Agreement of driving examiners' assessments – Evaluating the reliability of the Swedish driving test. *Transportation Research Part F: Traffic Psychology and Behaviour*, *19*(0), 22-30. doi: http://dx.doi.org/10.1016/j.trf.2013.02.004

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Baughan, C. J., & Simpson, H. (1999). Consistency of driving performance at the time of the L-test, and implications for driver testing. In G. B. Grayson (Ed.), *Behavioural Research in Road Safety IX*. Crowthorne: Transport Research Laboratory.

Berk, R. A. (1980). A Consumers' Guide to Criterion-Referenced Test Reliability. *Journal of Educational Measurement, 17*(4), 323-349. doi: 10.1111/j.1745-3984.1980.tb00835.x

Brennan, R. L. (2004). *Manual for BB-CLASS: A Computer Program that uses the Beta-Binomial Model for Classification Consistency and Accuracy. Version 1.* (CASMA Research Report No. 9). Retrieved from the Center for Advanced Studies in Measurement and Assessment at The University of Iowa website: http://www.education.uiowa.edu/docs/default-source/casma---research/09casmareport.pdf?sfvrsn=2

Brennan, R. L. (Ed.) (2006). *Educational measurement*. (4th ed.) Westport, CT: Praeger Publishers.

Brennan, R. L., & Wan, L. (2004). *Bootstrap procedures for estimating decision consistency for single-administration complex assessments* (CASMA Research Report No. 7). Iowa City: University of Iowa, Center for Advanced Studies in Measurement and Assessment. Retrieved from http://www.education.uiowa.edu/centers/casma/publications-data-file

Breyer, F. J., & Lewis, C. (1994). *Pass-Fail Reliability for Tests with Cut-Scores: A Simplified Method. ETS Research Report Series*, 1994(2), i-30. doi: 10.1002/j.2333-8504.1994.tb01612.x

Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. New York, NY: Holt, Rinehart and Winston, Inc.

Guo, F. (2006). Expected classification accuracy using the latent distribution. Practical Assessment, *Research & Evaluation*, *11*(6), 1-6. Retrieved from http://pareonline.net/getvn.asp?v=11&n=6

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15*(3), 309-333. doi: 10.1207/S15324818AME1503_5

Han, K. T., & Hambleton, R. K. (2007). *User's Manual: WinGen* (Center for Educational Assessment Report No. 642). Amherst, MA: University of Massachusetts, School of Education. Retrieved from http://www.umass.edu/remp/software/simcata/wingen/homeF.html

Hambleton, R. K., Swaminathan, H., Algina, J., & Coulson, D. B. (1978). Criterion-referenced testing and measurement: A review of technical issues and developments. *Review of Educational Research*, 1-47. Retrieved from http://www.jstor.org/stable/1169908

Hanson, B. A., & Brennan, R. L. (1990). An Investigation of Classification Consistency Indexes Estimated under Alternative Strong True Score Models. *Journal of Educational Measurement, 27*(4), 345-359. doi: 10.1111/j.1745-3984.1990.tb00753.x

Henriksson, W., Sundström, A., & Wiberg, M. (2004). *The Swedish driving-license test: A summary of studies from the department of educational measurement.* (EM 44) Umeå: Department of Educational Measurement, Umeå University. Available from the Umeå university website: http://www.jus.umu.se/digitalAssets/59/59522_em-45.pdf

Huynh, H. (1976). On the reliability of decisions in domain-referenced testing. *Journal of Educational Measurement, 13*(4), 253-264. doi: 10.1111/j.1745-3984.1976.tb00016.x

Huynh, H. (1990). Computation and statistical inference for decision consistency indexes based on the Rasch model. *Journal of Educational and Behavioral Statistics, 15*(4), 353-368. doi: 10.3102/10769986015004353

Huynh, H., & Saunders, J. C. (1980). Accuracy of Two Procedures for Estimating Reliability of Mastery Tests. *Journal of Educational Measurement, 17*(4), 351-358. doi: 10.2307/1434874

Lathrop, Q. N. (2015). Practical Issues in Estimating Classification Accuracy and Consistency with R Package cacIRT. *Practical Assessment, Research & Evaluation, 20*(18), 2. Retrieved from http://pareonline.net/getvn.asp?v=20&n=18

Lee, W. C. (2010). Classification Consistency and Accuracy for Complex Assessments Using Item Response Theory. *Journal of Educational Measurement, 47*(1), 1-17. doi: 10.1111/j.1745-3984.2009.00096.x

Lee, W.-C., Hanson, B. A., & Brennan, R. L. (2002). Estimating consistency and accuracy indices for multiple classifications. *Applied Psychological Measurement, 26*(4), 412-432. doi:10.1177/014662102237797

Livingston, S. A., & Lewis, C. (1995). Estimating the Consistency and Accuracy of Classifications Based on Test Scores. *Journal of Educational Measurement, 32*(2), 179–197. Retrieved from http://www.jstor.org/stable/1435147

Meyer, J. P. (2010). *Understanding Measurement: Reliability*. New York: Oxford University Press.

Peng, C. Y. J., & Subkoviak, M. J. (1980). A Note on Huynh's Normal Approximation Procedure for Estimating Criterion-Referenced Reliability. *Journal of Educational Measurement, 17*(4), 359-368. doi: 10.1111/j.1745-3984.1980.tb00837.x

Reiner, T. W., & Hagge, R. A. (2006). *Evaluation of the class C driver license written knowledge tests*. Retrieved from the State of California Department of Motor Vehicles website: http://www.dmv.ca.gov/portal/wcm/connect/b01cf8b0-d6e4-4532-86f6-3a0ddb791542/S2-221.pdf?MOD=AJPERES&CONVERT_TO=url&CACHEID=b01cf8b0-d6e4-4532-86f6-3a0ddb791542

Rudner, L. M. (2005). Expected classification accuracy. *Practical Assessment Research & Evaluation*, 10(13). Available online: http://pareonline.net/getvn.asp?v=10&n=13

Schuwirth, L. W. T., & van der Vleuten, C. P. M. (2011). General overview of the theories used in assessment: AMEE Guide No. 57. *Medical Teacher, 33*(10), 783-797. doi: 10.3109/0142159X.2011.611022

Siegrist, S. (Ed.). (1999). *Driver training, testing and licensing - towards a theory-based management of young drivers' injury risk in road traffic. Results of EU-project GADGET, Work Package 3*. BFU-report 40. Bern: Schweizerische Beratungsstelle Für Unfallverhütung.

Subkoviak, M. J. (1976). Estimating reliability from a single administration of a criterion-referenced test. *Journal of Educational Measurement, 13*(4), 265-276. doi: 10.1111/j.1745-3984.1976.tb00017.x

Subkoviak, M. J. (1978). Empirical Investigation of Procedures for Estimating Reliability for Mastery Tests. *Journal of Educational Measurement, 15*(2), 111-116. doi: 10.2307/1433864

Subkoviak, M. J. (1988). A practitioner's guide to computation and interpretation of reliability indices for mastery tests. *Journal of Educational Measurement, 25*(1), 47-55. doi: 10.1111/j.1745-3984.1988.tb00290.x

Sundström, A. (2003). *Den svenska förarprövningen. Sambandet mellan kunskapsprovet och körprovet, provens struktur samt körkortsutbildningens betydelse [Driver testing in Sweden. A study of the relationship between the theoretical and practical test, the structure of the tests and the effect of driver education on test performance]*. (PM 183). Umeå: Pedagogiska institutionen, enheten för pedagogiska mätningar. Available from the Academic Archive On-line DiVA website: http://umu.diva-portal.org/smash/get/diva2:588958/FULLTEXT01.pdf

Wang, T., Kolen, M. J., & Harris, D. J. (2000). Psychometric properties of scale scores and performance levels for performance assessments using polytomous IRT. *Journal of Educational Measurement, 37,* 141-162. doi: 10.1111/j.1745-3984.2000.tb01080.x

Wainer, H. (Ed.). (2000). *Computerized adaptive testing: A primer*. (2nd Edition). Mahwah, NJ: Lawrence Erlbaum Associates.

van der Linden, W. J., & Glas, C. A. (2010). *Elements of adaptive testing*. New York, NY: Springer New York.

Wiberg, M. (2004). *Klassisk och modern testteori. Analys av det teoretiska och praktiska körkortsprovet [Classical and modern test theory: analysis of the theoretical and practical driving-license test]*. (BVM 5) Umeå universitet: Institutionen för beteendevetenskapliga mätningar. Available from the Academic Archive On-line DiVA website: http://umu.diva-portal.org/smash/get/diva2:467117/FULLTEXT01.pdf

Wiberg, M., & Sundström, A. (2009). A comparison of two approaches to correction of restriction of range in correlation analysis. *Practical Assessment, Research & Evaluation, 14*(5), 2. http://www.pareonline.net/getvn.asp?v=14&n=5

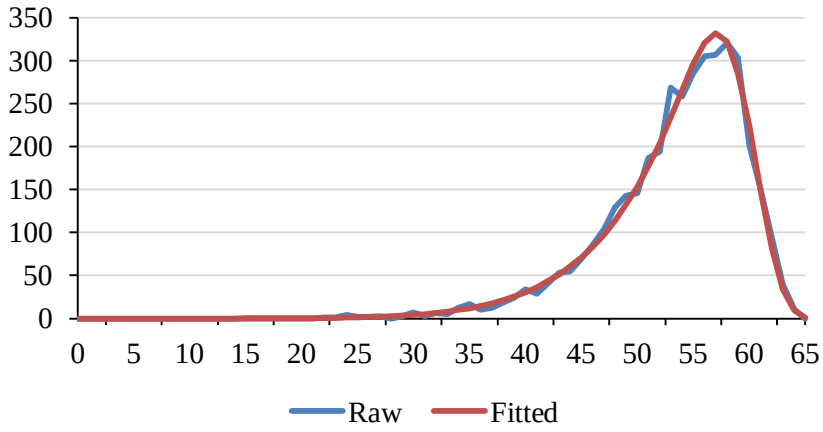
Woodruff, D. J., & Sawyer, R. L. (1989). Estimating Measures of Pass-Fail Reliability From Parallel Half-Tests. *Applied Psychological Measurement, 13*(1), 33-43. doi: 10.1177/014662168901300104

Wyse, A. E., & Hao, S. (2012). An Evaluation of Item Response Theory Classification Accuracy and Consistency Indices. *Applied Psychological Measurement, 36*(7), 602-624. doi: 10.1177/0146621612451522

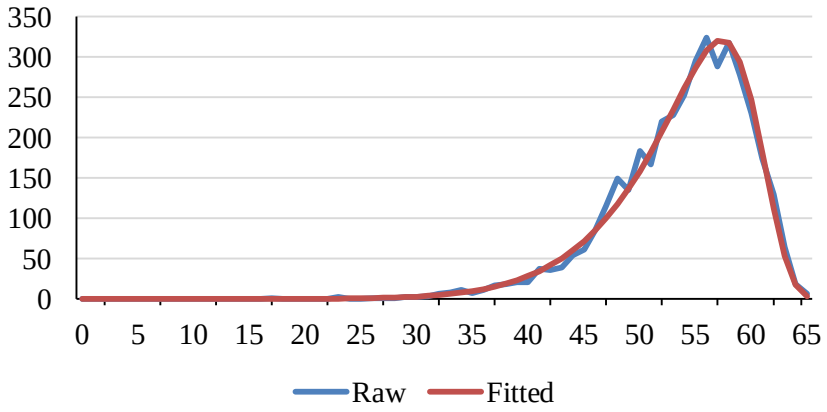
Appendix 1.

Graphs depicting model fit for the Hanson and Brennan 4-parameter model in terms of raw score and fitted score as frequencies at different score levels

Version 1



Version 2



Version 3