**Abstract**

The Icelandic National examinations in the Icelandic language is a test that assesses two content domains (e.g., reading and language). The fact is that nowadays, testing programs are facing increasing demands for reporting subscores in addition to the total score because reporting subscores could provide much-needed information about proficiency in the assessment for test-takers, teachers, and administrators. This study explored three different methods of assigning subscores to the Icelandic National examinations in the Icelandic language for 4th-grade students in compulsory schools. The methods are Haberman's method of added-value which is based on the Classical test theory, Multidimensional Item response theory & Confirmatory factor analysis. All the data analyses were done by using R-studio software.

*Keywords:* Subscore, Subscale, Classical Test Theory (CFA), added-value, Confirmatory Factor Analysis (CFA), Multidimensional Item Response Theory (MIRT), The Icelandic National Examinations.

This article focuses on examining if the Icelandic National Examinations in the Icelandic language for 4th graders in compulsory schools has the psychometric quality to report subscores to test takers and other officials in Iceland's academic environment. The structure is as following; first we will analyze the Icelandic National Examinations in the Icelandic language for 4th graders, the standards and structure. Secondly, we will go over the psychometric quality of subscore in educational testing. Third, we will go over three different methods that have been used to report subscores and are in the framework of Factor Analysis, Classical test theory, and Item response theory. For the data work, we will only use packages within the R-software (R core team, 2020) framework. We are going to apply these three different methods to the data to see if there is any difference in conclusions or results. It should be noted that in this article the term “subscale” is to denote one construct (domain?) and the term “subscore” is referring to a score on a subscale.

**The Icelandic National Examinations – in the Icelandic language**

According to the 39th Act of law no. 91/2008 for compulsory schools in Iceland, The Ministry of education, science and culture should provide standardized examinations for all compulsory schools in Iceland. The Ministry has entrusted the Directorate of Education to handle an assessment submitted for all students in 4.- 7.- and 9.th grade in all compulsory schools. All students in the 4th and 7th grades should take the Icelandic National Examinations in the Icelandic language and Mathematics, and 9th grade should take the exams in the Icelandic language, Mathematics, and English language (Einarsdóttir, Einarsdóttir and Skúlason, 2019). The Icelandic National Examinations will here on be called the INE.

The INE structure is according to the Table of specifications (or Test specifications) where the articulation of the knowledge, ability, and skills are being measured (Perie and Huff, 2015; Schmeiser and Welch, 2006). Using the Table of specifications makes it possible to make different types of the same test that are still measuring the same concept even though they do not include the same items on each test (Kolen and Brennan, 2014). Also, using the Table of specifications method, items on the test are explicitly divided into different groups. Each group is measuring a type of category, making it possible to measure subscores on the INE. For more information about the INE, see the article by Einarsdóttir, Einarsdóttir, and Skúlason (2019) and Einarsdóttir and Einarsdóttir (2018).

The INE in the Icelandic language for 4th graders is submitted in the fall of every year. It is divided into two categories, reading which is 70% of the total exam score, and language, which is the last 30% of the total score (Einarsdóttir, Einarsdóttir, and Skúlason, 2019). According to the National curriculum guide, the INE in the Icelandic language is supposed to assess students' ability and skills to understand the concept of a text that they read; and due to the answers, it should be possible to evaluate their knowledge, ability, and skills to use the Icelandic language (Pétursson, Úlfsdóttir and Skúlason, 2019).

**Subscale for reading**  **Number of items on subscales**.

**n = 15**

**n = 16**

**n = 5**

**n = 4**

**Figure 1.** Test scoring.

The four categories that all are measuring different types of reading and are as followed:

*Literal comprehension* – is when the information’s appear directly in the text.

*Inferential comprehension* - are hidden information in text that the reader needs to find out what is being told in the text without it being written directly.

*Understanding* - of a single word that are used in the text, or the meaning of idioms or a phrase.

*Summarizing questions* - where there is being tried to get the overall understanding of the concept of the text or what the authors is trying to tell through his story (Einarsdóttir, Einarsdóttir and Skúlason, 2019).

**Psychometric Quality of Subscore**

National Curriculum Guide (2014) in Iceland notes when there is an evaluation of test scores it should be reliable, impartial, honest, and fair for test-takers where all aspects of education should be met by the criteria of the National Curriculum Guide. Subscores are test scores that often are evaluated to students or other educational authorities for better information on test scores. The simplest way to describe subscores would be to say that it is a cluster of items on a test that are measuring the same category (Sinharay, Puhan, Haberman, Hambleton, 2018). In some cases, subscales have the possibility to give more information than the total score and for that reason there is more often increasing demand from test users who want the subscores to be reported along with the total score because of their potential diagnostic value (Haberman and Sinharay, 2010; Brennan, 2012). This is also true for the Icelandic National examinations, where it has been increasing interest in reporting subscore resolute to test-takers. Nevertheless, it is very common for subscales on educational test to have very low reliability due to how few items are on each subscale (Haberman, 2008; Haberman, Sinharay & Puhan, 2009; Sinharay Puhan & Haberman, 2011) and may often be the outcome of a retrofitting, because usually the test where not designed to report subscores to begin with (Goodman and Hambelton, 2004).

Before the subscore is reported, they need to satisfy specific quality standards, Haberman (2008) noted that for subscores to be reported, it needed to have high reliability and be distinct from other subscores measured in the test. The Standards for Educational and Psychological Testing (2014) which includes relevant information suitable for all types of subscores, describes subscore often to be relatively small numbers of items that usually are less reliable than the total test score and might even being reflecting a measurement error, for that reason the decision to provide subscore results should be carefully made. The standard 1.14 notes, "When interpretation of subscores ... is suggested, the rationale and relevant evidence in support of such interpretation should be provided.". Standard 2.3 states that when subscores and other test scores are being reported to test-takers, those responsible for the testing programs should report to the user’s: reliability data for all test-scores in the test. It should be reported in a way so the test taker can make his own judgment of whether he wants to take advantage of the results or not.

**Factor analysis - CFA**

Method like Factor analysis is often used in educational research to determine the structure of the test. In the Factor analysis framework, both exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) are available when evaluating subscores. When using CFA, the subscore construct needs to be in composition (Puhan and Liang, 2011). The subscores are estimated from each subscale to determine each subscales dimensionality and see if they are distinct enough to be reported.

**Haberman's method of added value**

Haberman's (2008) method of added-value has been very noticeable in recent years where it is trying to come to a decision whether subscores should be reported or not. Haberman's method is based on the Classical Test Theory (Lord & Novick, 1968), where he examined if the subscore has added-value over the total score. Haberman suggested a value in reporting subscore if the subscore reliability is a more significant and more reliable predictor than the PRMSE of the total score. The use of the proportional reduction in mean squared error (PRMSE) is as followed; the PRMSE for Ss, Sx, and Ssx will be denoted as PRMSEs PRMSEx and PRMSEsx, respectively. The PRMSEs are the same number as the subscore reliability and therefore lies between 0 and 1. When the reliability is closer to one it indicates higher reliability, respectively. Where there is higher reliability it should mean lower mean square error. The PRMSEx is the reliability of the observed total score and a measure of the subtest construct. For the subscore to be reported, the PRMSEs need to be greater than the PRMSEx. for more information about the method, it can be found in the Haberman's (2008) article.

For augmented subscore to have added value the PRMSEsx needs to be larger than both PRMSEs and PRMSEx. Haberman and Sinharay (2013) suggest that the distance of PRMSEs or PRMSEx (whichever is larger) should be at least 10% for the augmented subscore to have added value and the following condition should be met:

PRMSEsx – (PRMSEs or PRMSEx) > 0.1 \* (1 – (PRMSEs or PRMSEx)).

Like noted above this method has been noticeable in recent years and applications of the method can be found in many articles like Sinharay, Puhan, and Haberman (2010), Haberman and Sinharay (2013), Lyren (2009), Sinharay (2010), Meijer, Boev, Tendeiro, Bosker and Albers (2017), Puhan, Sinharay, Haberman and Larkin (2010), Sinharay, Puhan and Haberman (2011)

Sinharay, Haberman and Puhan (2007) used Haberman's method on two types of basic skills test where the results did not provide any more value in reporting subscores to examinees or institutions than the total score.

Sinharay (2010) he took the Haberman’s method and simulated data from a several operational datasets where subscore were founded to have added value over total score. With that data he also examined what properties subscore should have to be consider having added value. His conclusions where that for subscore to have added value over the total score, it needs to consist at least 20 items in each subscale and to they need to be sufficiently distinct from each other.

Lyrén (2009) applied the method of added value to the college admission test (SweSAT) where his result showed that there is a value in reporting the subtest score for three of five subtest that he was examining (Vocabulary; 40 items, Data Sufficiency; 22 items & Diagram, Tables and Maps; 20 items). The two other subtest score that was being examined were Swedish Reading Comprehension (20 items) and English Reading Comprehension (20 items) and it did not indicate that they were with added value. Lyrén noted in his results that the Swedish Reading was thou with good reliability, and the true reading score were good measure of the score.

**MIRT - Multidimensional Item Response Theory**

With Item response theory you are analyzing the probability of a person answering certain trait or ability correctly due to the item difficulty and discrimination (DeMars, 2010). When Multidimensional Item Response theory (MIRT) (Reckase, 1997) model is applied to a data it needs to be measuring more than one ability or skill, and those multiple skills or ability’s needs to be measuring one construct (domain) each. When using MIRT models to measure subscores then one subscale is one construct (Rackase, 2009) where it calculates information across subscores to be able to improve the reliability estimates, where it borrows information from another subscore in the same subgroup from the persons own test scores (Longbach & Payton, 2017).

Several researchers have examined the use of MIRT models to report subscores, like Haberman and Sinharay (2010) applied the MIRT method to several operational datasets to find out if they have added value over the total score. They found out that the MIRT method was slightly more accurate than CTT-based methods, and more likely to have added value. Longbach and Peyton (2017) did a research where they tried different methods to assigning subscores on the K-12 English language proficiency tests, where they found the MIRT method to be the most reliable method to examen if subscores have the psychometric quality to be reported.

purpose of

factor analysis is to discover simple patterns in the pattern

of relationships among several observed variables. In partic-

ular, it seeks to discover if the observed variables can be

explained largely or entirely in terms of a small number of

variables called factors. A simple factor-analytic approach to

evaluate whether the subscores are distinct enough would be

to compute the eigenvalues from the correlation matrix of

the subscores (or from the correlation matrix of the items).

If most of the eigenvalues computed from the correlation ma-

trix of the subscores are smaller than 1 or if a scree plot of

these eigenvalues shows that the eigenvalues abruptly levels

out at some point, then the number of factors in the data is

less than the number of subscores and the claim of several

distinct subscores is probably not justiﬁed. The presence of

multiple factors, that is, multiple large eigenvalues, would

support the reporting of subscores.

purpose of

factor analysis is to discover simple patterns in the pattern

of relationships among several observed variables. In partic-

ular, it seeks to discover if the observed variables can be

explained largely or entirely in terms of a small number of

variables called factors. A simple factor-analytic approach to

evaluate whether the subscores are distinct enough would be

to compute the eigenvalues from the correlation matrix of

the subscores (or from the correlation matrix of the items).

If most of the eigenvalues computed from the correlation ma-

trix of the subscores are smaller than 1 or if a scree plot of

these eigenvalues shows that the eigenvalues abruptly levels

out at some point, then the number of factors in the data is

less than the number of subscores and the claim of several

distinct subscores is probably not justiﬁed. The presence of

multiple factors, that is, multiple large eigenvalues, would

support the reporting of subscores.

purpose of

factor analysis is to discover simple patterns in the pattern

of relationships among several observed variables. In partic-

ular, it seeks to discover if the observed variables can be

explained largely or entirely in terms of a small number of

variables called factors. A simple factor-analytic approach to

evaluate whether the subscores are distinct enough would be

to compute the eigenvalues from the correlation matrix of

the subscores (or from the correlation matrix of the items).

If most of the eigenvalues computed from the correlation ma-

trix of the subscores are smaller than 1 or if a scree plot of

these eigenvalues shows that the eigenvalues abruptly levels

out at some point, then the number of factors in the data is

less than the number of subscores and the claim of several

distinct subscores is probably not justiﬁed. The presence of

multiple factors, that is, multiple large eigenvalues, would

support the reporting of subscores.

purpose of

factor analysis is to discover simple patterns in the pattern

of relationships among several observed variables. In partic-

ular, it seeks to discover if the observed variables can be

explained largely or entirely in terms of a small number of

variables called factors. A simple factor-analytic approach to

evaluate whether the subscores are distinct enough would be

to compute the eigenvalues from the correlation matrix of

the subscores (or from the correlation matrix of the items).

If most of the eigenvalues computed from the correlation ma-

trix of the subscores are smaller than 1 or if a scree plot of

these eigenvalues shows that the eigenvalues abruptly levels

out at some point, then the number of factors in the data is

less than the number of subscores and the claim of several

distinct subscores is probably not justiﬁed. The presence of

multiple factors, that is, multiple large eigenvalues, would

support the reporting of subscores.

purpose of

factor analysis is to discover simple patterns in the pattern

of relationships among several observed variables. In partic-

ular, it seeks to discover if the observed variables can be

explained largely or entirely in terms of a small number of

variables called factors. A simple factor-analytic approach to

evaluate whether the subscores are distinct enough would be

to compute the eigenvalues from the correlation matrix of

the subscores (or from the correlation matrix of the items).

If most of the eigenvalues computed from the correlation ma-

trix of the subscores are smaller than 1 or if a scree plot of

these eigenvalues shows that the eigenvalues abruptly levels

out at some point, then the number of factors in the data is

less than the number of subscores and the claim of several

distinct subscores is probably not justiﬁed. The presence of

multiple factors, that is, multiple large eigenvalues, would

support the reporting of subscores.

**Method**

*Participants*

The participants for this study were 11,027 4th graders in Icelandic compulsory schools who took the INE assessment in fall 2018 (*n* = 4457), fall 2019 (*n* = 4462), and fall 2020 (*n* = 2108).

*Test Data*

The INE assessment is applied online, and students have access to a computer or pad in their school to submit the exam. The test has only multiple-choice (MC) items response where each question includes four possibilities and one correct answer. For this application, we will only be testing if the reading part has the psychometric quality to measure subscores. The INE is divided into four types of reading (40 items) witch composites four subscales scores, LEB (Literal comprehension, 15 items), LEC (Inferential comprehension, 15 items), and LED (Understanding, 4 items), LEE (Summarizing questions, 5 items). The INE was not designed to give subscores only a total score of the two factors reading and language, but due to the INE structure with a Table of specifications, it has that possibility.

*Data analysis and Software*

The INE was scored using Haberman’s CTT theory, CFA, and the MIRT method; reliability was estimated for each method. For the data analysis were used three different packages within the framework of R-software (R core team, 2020).

*CFA*

For the CFA were used the “lavaan” (Rosseel, 2017) package, to estimate subscores as factor scores for each subscale. Using the lavaan package the recommendations about if you should report subscore or not are not automatically given, so the subscore psychometric quality needs to be evaluated by looking at the model fit indexes. The internal consistency was estimated with Cronbach’s alpha coefficient on each subscale, where the criterion for good reliability was set for higher than a = .80 for respectable reliability (DeVellis, 2016). It is recemented to use more than one fit index when evaluating model fit statistics and in this study were used the Comparative fit Index (CFI), the Tucker-Lewis Index (TLI), and the Root mean square error of approximations (RMSEA). CFI indicates if the model fits the data better than a model that does not assume a relationship between the measured variables. The CFI fit is in the range from zero to one, where a higher number indicates a better fit to the model, and fit higher than .90 would indicate a good fit, but fit higher than .95 would mean the data fits the model very well (Raykov and Marcoulides, 2006). TLI is similar to CFI, where it measures how well the model fits the data. TLI ranges from zero to one where a good model fit would be .95 (Brown, 1985; West, Taylor and Wu, 2014). The RMSEA index considers the model complexity where it indicates the degree of discrepancy between the model and the data. For a good RMSEA fit, the value needs to be less than .05 for a reasonable approximation (Browne and Cudeck, 1993; Raykov and Marcoulides, 2006).

*MIRT*

For the MIRT analysis where used the “mirt” (Chalmers, 2020) package in R to obtain subscore for each subscale in the framework of MIRT. When it comes to the decision on whether subscore in the MIRT framework, should be reported or not it should be made by evaluating the model-data fit statistics and see if they are of an acceptable level. The Model fit indexes that are available in the mirt package are as previously noted for the CFA model, CFT, TLI and RMSEA.

*CTT*

For the Haberman’s method were used the R package “subscore” (Dai, Wang and Svetina, 2016) to examine if they are of added value. To do that there are two ways (1) the PRMSEs needs to be grader than PRMSEx or (2) for the augmented subscores you can report subscores if the PRMSEsx is at least 10% larger then ether PRMSEx or PRMSEs.

There have been practitioners that have believed that subscores with few items on the subscale may have added value if the are sufficiently distinct from each other, how ever Sinharay (2010) study

Sinharay (2010) did a simulation based on Haberman’s method on a data where he said for subscore to have added value over the total score the subscale needs to consist of more than 20 items. Lyrén (2015) did a sampling on his data to find recommended sample size for Haberman’s method and found out that it was required a sample size of at least 1200 participants. Lyrén (2008) And he also examined what test should need to possess in order to report subscore

**Results**

Skrifa texta hér um hvernig niðurstöðuarnar eru birtar.

**Table x.** Correlations between subscales within the three methods.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Method | LEB | LEC | LED | LEE |
| Haberman’s | .693 | .662 | .415 | .340 |
| CFA | .039 | .038 | .016 | .020 |
| MIRT |  |  |  |  |

*Haberman’s method*

The results for Haberman’s method

Table x.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Alpha | PRMSE.s | PRMSE.x | PRMSE.sx | Added value.s | Added value.sx |
| LEB | .693 | .693 | .861 | .862 | - | - |
| LEC | .662 | .662 | .832 | .832 | - | - |
| LED | .415 | .415 | .788 | .791 | - | - |
| LEE | .340 | .340 | .854 | .854 | - | - |
| Total | .843 |  |  |  |  |  |

*Confirmatory Factor Analysis*

*df* = 734, *p*<.001, and other fit are, Comparative Fit index (CFI) is .894, Tucker-Lewis Index (TLI) is .888, Root mean Square Error of Approximation (RMSEA) is .025

**Conclusions**

Data requirements

**References**