Model Selection using a Stepwise Bayesian Information Approach in Multiple Group Models with Binary Data

Nasseem Hessami

Master's Programme
Assessment, Measurement and Evaluation
120 credits

University of Oslo
Faculty of Educational Sciences

May 15, 2020

Popular Abstract

High stakes tests have the power to impact almost every facet of the test taker's life. Whether a student proceeds to the next grade level, or a lawyer is permitted to practice law, can and are, on a daily basis shaped by the results on standardized tests. An important complication arises, though, when it is considered that not all tests are created equal, and what may be considered an easily interpretable question for a test taker with certain group membership—say, gender or ethnicity or socioeconomic status—might prove quite difficult for another test taker, simply because they come from a different background. This experiment explores a model selection method that might succeed in helping craft and analyze fairer tests, so test takers are assessed on the skill being measured, not irrelevant characteristics or qualities. The results of this experiment seem to suggest that the model selection method explored was successful, and test agencies may wish to implement this approach for future test design and analysis procedures.

Acknowledgements

I would like to thank the faculty and staff at the University of Oslo's Center for Educational Measurement, for their unwavering support and dedication throughout the entirety of this program. In particular, I extend my deepest gratitude to my supervisor, Dr. Björn Andersson, for his continual guidance, innovation, patience and mentorship throughout this endeavor. Thank you for making my daydreams of reaching this point, a reality.

Finally, I thank my beloved family and friends for everything, with special emphasis on my father, Nasser Hessami. He was, and is, education's biggest advocate.

Abstract

High stakes tests are the focus of heightened attention as they continue to influence the futures and opportunities of test-takers, but design flaws can lead to tests operating differently for members of certain groups. In the context of item response theory (IRT), the Bayesian Information Criterion (BIC) is capable of isolating the source of non-invariance in a test or model due to its consistency property when the sample size is sufficiently large and the true model exists in the candidate pool. As its theory currently stands, though, the BIC is not of use to most non-invariance investigations due to its inability to efficiently move through enormous candidate model pools. Using the BIC in a novel stepwise approach, results from this simulation experiment support that the novel stepwise approach can isolate the true model with the exact accuracy of the traditional BIC procedure in a fraction of the time, given a sufficiently large sample size. The initial results of this experiment suggest that, contingent on further research and investigation, the stepwise approach may be of use to high stakes test agencies to ensure fair testing practices.

Keywords: IRT, BIC, DIF, Heuristic, Stepwise, Model Selection, 2PL

Model Selection using a Stepwise Bayesian Information Approach in Multiple Group Models with Binary Data

On a global scale, standardized tests and high stakes exams have received increasing praise and criticism in recent years, as they have impactful consequences on a local level, for example determining if a pupil advances to the next level of education or secures a prestigious job, as well as impactful consequences on a macro level, sometimes shaping the entire education policy implemented by a state or nation (Froese-Germain, 2001). Considering the vast array of ways in which test takers' futures and opportunities are shaped by the tests administered on a local, national, and international level, it is not difficult to understand the public outcry and controversy that results when high stakes tests are sometimes revealed as flawed or errored in either their content or implementation (Newton, 2005). One of many possibilities for a test to reveal itself as flawed is when it is discovered that different test-takers interpret one or more of its items differently, due to traits or characteristics that do not pertain to the ability the item or test intends to measure. For example, an arithmetic problem that describes fluctuating rain patterns during monsoon season may be interpreted correctly by test takers from coastal, tropical regions, but incorrectly by those from desert-like regions marked by long droughts.

As high stakes tests have the potential to severely influence test takers' personal and academic opportunities, it is necessary for procedures and tools that can intercept and isolate biased items to continue to be built out, developed, and fine-tuned. In this way, tests are part of a continual endeavor undertaken by test agencies and designers to operate in as fair a manner as possible, so that test-takers' odds of answering items correctly depends solely on their possession of the skill or ability the item intends to measure in the first place. When tests are designed in accordance with a school of thought called Item Response Theory (IRT), there are a variety of ways in which biased items can be isolated or targeted. One of those ways involves model selection using the Bayesian Information Criterion (BIC) developed by Schwarz (1978). When the item count of a test suspected of having biased items is relatively high, however, the process for uncovering biased items becomes more complicated, due to the exponentially increasing number of combinations in which biased items might reveal themselves. Resultantly, by developing a procedure in which far fewer combinations of potentially biased items need be assessed, it was the aim of this paper to explore if a more efficient, faster way to identify biased items in real-world test settings was possible. This stepwise BIC procedure was tested across a variety of different settings and conditions to determine its utility and applicability to real-world investigations that seek to isolate and prevent bias in high stakes tests.

As such, a brief explanation of the themes and schools of thought pertinent to the experiment are discussed in this section. First, a brief summary of IRT, the theory for which the Two Parameter Logistic (2PL) models used in this experiment adhere to, is given. The mathematical formula of the 2PL model, as well as its application in modern testing is also discussed. Next, the model selection process and different criteria, namely, the Bayesian Information Criterion (BIC) is summarized. Finally, Differential Item Functioning (DIF) is explained in context of IRT, highlighting the ways in which DIF complicates the model selection process, in particular when using the Bayesian Information Criterion.

Item Response Theory (IRT)

In educational and psychological settings, tests serve as instruments that can measure or gauge the extent to which test-takers possess a certain skill, behavior, or trait (Kelderman & Rijkes, 1994). Within the last century, different schools of thought concerning the prediction and explanation of test scores have emerged. Item Response Theory (IRT) is one of the prominent test theories used in educational and psychometric settings today, as its advantages over the traditionally used classical test theory are considered well documented in the scientific community (Nguyen, Han, Kim, & Chan, 2014). IRT is conceptualized as a set of statistical models that explain the connection between an individual's abilities and how he or she responds to questions on a scale (Nguyen et al.,2014). While not a theory, strictly speaking, proper implementation of IRT allows for test developers and agencies to predict and explain anticipated responses, and consequently scores, when administering a test to a population that possesses varied levels of a particular trait the test is meant to measure. Proper implementation of IRT, however, is contingent upon a specific set of assumptions. That is, to credibly predict and explain scores for a given test, the following assumptions, per Ayala (2009), must be met:

1. Unidimensionality of the trait that the test intends to measure. That is, outside of rarer cases when the test seeks to measure multiple traits simultaneously, the test should measure a clear, singular trait or quality in the population.
2. Local independence across participants and their responses to items. That is, participants' responses to a given item on the test is determined solely by participants' possession of the trait in question, and no other influences.
3. Functional form of the response data. That is, based on the test's item properties, the correct IRT model is selected, and test response data follows the distribution and shape dictated by that model.

Assuming the necessary assumptions are satisfied, IRT item parameters should, theoretically, perform invariantly across different populations, even when the populations possess different levels of the trait in question (Ayala, 2009). Population invariance is defined as the phenomenon in which item parameters estimated by an IRT model remain constant across different populations of test takers (Nguyen et al., 2014). Resultantly, this allows for predictions and conclusions pertaining to one population's performance on a test to be applied to numerous populations, a feature not possible in other classical approaches.

The Two Parameter Logistic (2PL) Model

Per the aforementioned IRT assumptions, the models comprising IRT are varied and diverse, and selecting the appropriate model for a particular population is based on the item properties of the test in question. Two Parameter Logistic (2PL) models correspond to tests scored in a binary fashion, whose items contain two parameters of interest: the discrimination (α) parameter and the difficulty (δ) parameter. The discrimination parameter pertains to an item's ability to differentiate between individuals with different ability levels (Nguyen et al., 2014), with reasonably good values ranging from 0.8 to 2.5 (Ayala, 2009). An item's difficulty parameter, meanwhile, corresponds to the point on the latent continuum where the test taker from a given population has a 50% probability of selecting the correct answer (Nguyen et al., 2014). The difficulty parameter typically ranges from –3 to 3 (Ayala, 2009). Lastly, the latent continuum refers to the distribution of theta (θ), the trait the test theoretically measures. Theta is a variable that refers not to the scale's items, but to the skill, ability, or proficiency level of the individual completing the scale (Nguyen et al., 2014). As theta is directly related to an item's difficulty parameter, values for theta often range from –3 to 3, as well (Ayala, 2009). Using Ayala's (2009) notation, the 2PL model is explained through the logistic formula:

$$p\left(x_j = 1 \middle| \theta, \alpha_j, \delta_j\right) = \frac{e^{\alpha_j(\theta-\delta_j)}}{1+e^{\alpha_j(\theta-\delta_j)}}\,, \tag{1}$$

where, x_j is the score on item j, θ is the latent trait possessed by the participant, α_j is the discrimination parameter of item *j*, and δ_j is the difficulty parameter of item *j*.

The IRT models implemented for different large-scale assessments like state level end-of-course tests are usually dictated by specific policies, and relevant policy makers usually strive for best absolute fit of the model relative to observed test responses (Brown, Templin, & Cohen, 2015). Many well-known high stakes exams and tests have been designed around Item Response Theory, specifically the 2PL model. For example, the Test of English as a Foreign Language (TOEFL) exam, as well as the Graduate Record Examination (GRE) are designed in accordance with the 2PL model (Ricarte, Cúri, & von Davier, 2017). Given its frequent implementation in international high stakes tests, model selection abilities of the 2PL model were the focus of the experiment described herein.

Model Selection

Models "provide a framework for reasoning about patterns across any number of unique real-world situations, in each case abstracting salient aspects of those situations and going beyond them in terms of mechanisms, causal relationships, or implications at different scales or time points that are not apparent on the surface" (Mislevy, 2009, p. 11). Often evaluated in the exploratory phase of data analysis, models are selected from a pool of candidate models using different criteria that take into account model fit and the number of parameters required to obtain such fit (Sclove, 1987).

The Bayesian Information Criterion (BIC). There are many criteria that guide model selection from a pool of candidate models, but one criterion – the Bayesian Information Criterion (BIC) developed by Schwarz (1978) – is of particular interest due to its consistency property. According to Vrieze (2012), the consistency property holds that as the sample size grows large, the BIC selects the true model with probability approaching 1, assuming its other statistical conditions are satisfied and there is a fixed number of candidate models under consideration.

Many variations of Schwarz's (1978) BIC are in circulation. The variation used by Kim, Cohen, Cho, and Cho (2019) demonstrates the penalty functions the BIC is built on:

$$BIC\ (\hat{\theta}) = \ -2\ logL(\hat{\boldsymbol{\theta}}) + k \log N, \tag{2}$$

where, $L(\hat{\boldsymbol{\theta}}) = p(\boldsymbol{y}|\hat{\boldsymbol{\theta}})$ is the likelihood function for which **θ** designates the set of parameters, its hat indicates the maximum likelihood estimates, $\boldsymbol{y}$ designates data, *k* is the number of parameters, and *N* is the sample size. It is important to note that *log N* in Equation 2 serves as the BIC's penalty function, which corrects for maximum likelihood estimation. Unlike other criterions, like the Akaike Information Criterion (AIC), the BIC's penalty function increases as the sample size increases, meaning that as the sample size grows, it is less and less likely to achieve statistical significance, meaning the BIC's Type I error rate approaches zero as N increases (Vrieze, 2012). Per Equation 2, assuming the BIC's necessary statistical conditions are satisfied, given a candidate model pool with a sufficiently large sample size, the model with the lowest BIC has the highest likelihood of being the true model, or, per Vrieze (2012), the model that originally generated the data.

Elements of Model Selection. If the true model is conceptualized as the model that initially generated the data, in context of a non-invariance investigation, the true model may or may not adhere to the invariance property in one or more of its items. Violation of the invariance property, or non-invariance, is not an unheard-of phenomenon when dealing with IRT models, and different procedures exist for test designers to isolate and troubleshoot non-invariance. One such way is to conduct a non-invariance investigation via a model selection procedure. In non-invariance investigations, as well as other model selection situations where the true model is among the pool of candidate models, a criterion like the BIC is appropriate and is tasked with moving through the candidate pool and selecting the best-performing model, or, model with the highest likelihood of being the true model. While theory pertaining to the BIC supports its tendency to select the true model when certain assumptions are satisfied, there is of course no guarantee that the selected model – the model selected from the candidate pool – is in fact the true model –the model that initially generated the data. In model selection procedures, it is possible that a criterion such as the BIC selects either the true model from the candidate pool, or a candidate model (see Figure 1). The odds of the true model being selected is impacted by a variety of factors, for example, how well the BIC's assumptions are satisfied.

The Model Selection Procedure

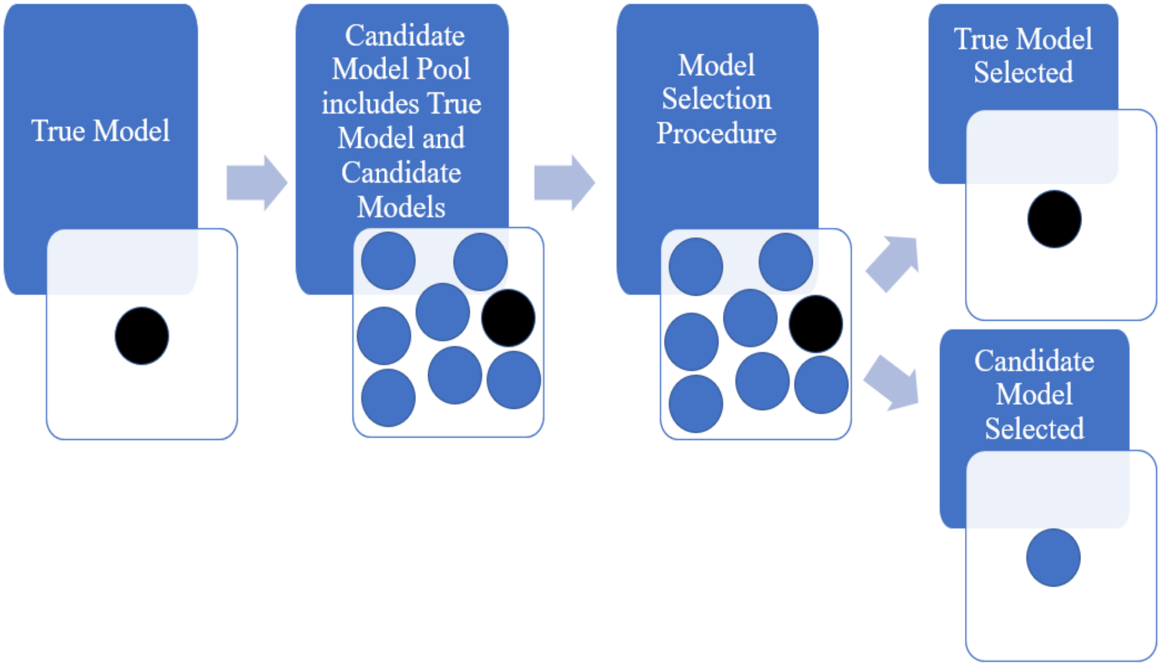
True Model
Candidate Model Pool includes True Model and Candidate Models
Model Selection Procedure
True Model Selected
Candidate Model Selected

Figure 1. In a model selection procedure, it is possible for the selection approach to select either the true model (black) or one of many non-true candidate models (blue) from the candidate pool.

Population Invariance and Differential Item Functioning (DIF)

The discrimination and difficulty parameters estimated in a 2PL model play a role not just in defining an instrument's items, but in a broader context underscore the ability to compare or predict different populations' responses to a given instrument. That is, when the conditions for an IRT-adhering test are upheld, in theory, one of the defining features achieved through the test is population invariance, that its items behave invariantly across different groups.

Even after controlling for differences in the latent trait, however, population invariance is not necessarily always achieved (Ayala, 2009). When population invariance is violated, one or more of the test's items function differently across different populations or subgroups. This phenomenon is known as differential item functioning (DIF), or non-invariance, and can occur because a test is not measuring the intended trait, is measuring unintended traits in addition to the intended trait, or is measuring the intended trait imprecisely, with excessive noise or error (Ayala, 2009).

When the population invariance condition is satisfied, the discrimination and difficulty parameters that are estimated across two groups are identical, or invariant. When the population invariance condition is violated, however, estimated parameters may reflect one set of values for one group, but a different set of values for the other. Therefore, in context of model selection, if the objective is to determine if a test exhibits DIF, all possible combinations of invariance and non-invariance across all items comprising the test must be explored. If model selection is guided by the best-fitting model and the BIC is used to assess model fit, as is the case in the experiment described herein, the selected model is that which corresponds to the lowest BIC in a candidate model pool. In the event the selected model has better fit, or a lower BIC, than the candidate model in which all item parameters are constrained to be equal, this would theoretically suggest that the true model likely exhibits DIF in some capacity.

All combinations of invariance and non-invariance must be assessed for each subgroup under investigation to determine which candidate model has the best fit, or lowest BIC. As the number of subgroups or the number of items increases, this drastically increases the number of candidate models to be assessed. For example, if an investigation pertains to two groups taking a 10-item test, then 2^{10} or 1,024 candidate models must be assessed for best fit. For three groups, the candidate model pool spikes to 3^{10} or 59,049 competing models. Thus, non-invariance investigations that involve individually assessing every candidate model's BIC to isolate the best fitting model, and thereby determine if a test contains DIF, are arduous endeavors requiring intensive manpower and resources. As such, it is the objective of this paper to determine if there is a computationally more efficient way for a criterion like the BIC to move through an unfeasibly high number of models in non-invariance investigations.

Uniform and nonuniform invariance. The 2PL model has two item parameters of interest, the discrimination and difficulty parameter. Consequently, there are two forms of DIF, uniform DIF and nonuniform DIF, that warrant discussion in relation to non-invariance investigations. As its name suggests, uniform DIF behaves in a way such that the distribution of one group's responses to the item, called the item response function (IRF) is always higher or lower than that of the other group, throughout the entire latent continuum (Ayala, 2009). An item might express uniform DIF across two groups, for example, if throughout the entire latent continuum, members of the reference group consistently require less of the skill being measured, compared to the focal group, to get the question correct. As far as the item parameters are concerned, uniform DIF therefore means different groups interpret the difficulty (δ) of an item differently, but the item's discrimination parameter (α) operates identically across groups. Conversely, when an item expresses nonuniform DIF, this means that the two groups' IRFs are not uniformly distributed in relation to one another throughout the latent continuum (Ayala, 2009). When an item expresses nonuniform DIF, for example, perhaps at one point in the latent continuum, the focal group requires less of the given trait to outscore the reference group, but at a different point in the latent continuum, the circumstances are reversed. In nonuniform DIF, when the IRFs are not identical across groups, this requires that either the item discriminates differently across groups (different α parameters in each group), or that both the discrimination (α) and difficulty (δ) parameters differ across groups.

Research Questions and Study Aims

Given the reliability and popularity of using the BIC in model assessment procedures when its conditions are satisfied, the aim of this experiment is to explore conditions where BIC-guided model selection have yet to be explored. Considering the complexity of using the criterion for unfeasibly large candidate pools, as is the case for non-invariance investigations, the research questions this paper aims to answer are:

1. Via a novel stepwise approach, starting from the most constrained model possible and sequentially relaxing parameters, does the BIC tend to select the true model with probability approaching 1 as sample size and/or number of items increase?
2. In either the traditional or novel stepwise BIC approach, does the concentration or magnitude of non-invariance impact the approach's ability to select the true model as sample size and/or number of items increase?
3. What is a sufficiently large sample size for the traditional BIC method to select the true model with probably approaching 1?

4. In the conditions where the true model is successfully isolated by the traditional approach, does the stepwise approach isolate the true model to the same extent?

Through a carefully crafted simulation experiment, it was the objective of this study to seek to answer these questions. Considering the assumptions and behaviors of the traditional BIC, developed by Schwarz (1978), are relatively well-known and applied regularly by statisticians and researchers alike, a simulation control study was designed so as to not only confirm the behavior of the traditional approach when its assumptions were met, but also to explore how a stepwise approach might fare in identical, comparable settings. As no theory existed on the behavior of the BIC in isolating the true model in a stepwise procedure, it was necessary to craft an experiment that allowed both the traditional approach and stepwise approach to be implemented on a comparable instrument. For this reason, both approaches were tested on a six-item instrument over a variety of comparable settings. For comparison purposes, the same true model was established in the data generation phase of each applicable approach for a given setting, and trends in results were thereafter compared. The last component of the experiment was more exploratory in its motivation, testing only the stepwise BIC approach on a larger 20-item instrument, an instrument too computationally demanding to efficiently test the traditional approach on.

The remaining sections of this paper are organized as follows. First, a more comprehensive and detailed explanation of the BIC approaches comprising the experiment are provided in the methods section. Thereafter, comparisons in performance between the traditional and stepwise procedure are supplied in the results sections. Performance of the stepwise procedure on the 20-item instrument is also reported. Finally, major takeaways and interesting findings, as well as experimental limitations and concluding thoughts, are offered in the discussion section.

Method

This section opens with an explanation of the competing mechanisms central to the experiment. First the methodology behind the traditional BIC mechanism, and thereafter the methodology of the novel stepwise approach, are explained. Then, the different simulation conditions specific to the experiment, including the different sample sizes (N), item counts (*i*), DIF concentrations, and magnitude settings, are described.

Model Selection Procedures using BIC

The BIC model selection simulations were executed via two different methodological approaches consisting of a traditional algorithm approach and a novel stepwise algorithm approach. In both procedures, the fully relaxed, non-identifiable model was always excluded from the candidate pool.

Traditional BIC Method

In the traditional BIC model selection approach, the BIC of every possible candidate model (e.g. every possible combination of non-invariance) is assessed, and per Equation 2, the model with the lowest BIC is the model that is selected. Accordingly, the traditional mechanism operates as follows for a test with *i* items:

1. The BIC for all candidate models (2^i combinations of non-invariance – 1) is recorded in no particular order
2. Upon assessing the BIC of all candidate models, the model with the lowest BIC is the selected model

Stepwise BIC Method

Unlike the traditional mechanism, which assesses the BICs of all candidate models in no particular order, stepwise approaches begin by analyzing a certain set of candidate models, sequentially adjusting parameter constraints, and moving to the next set of models based on performance of the first set of models. In particular, the stepwise approach relevant to this experiment, always begins by assessing the most constrained model possible (e.g. the model with the most invariant item parameters across the reference and focal group). From this point forward, the number of relaxed items, *b*, for the models that comprise a given candidate pool, occur in a series of sequential steps, dictated by specific conditions. The candidate pool for the first step of relaxations is always comprised of models for which $b \leq 1$. That is, for the first step of relaxations, the candidate pool includes the most constrained, completely invariant model ($b = 0$), and competitive models in which $b + 1$ items are relaxed. The BICs of the initial candidate pool's models are assessed, and if the lowest BIC corresponds to the most constrained model in the candidate model pool, it is the selected model. If the lowest BIC does not correspond to the most constrained model in the candidate group, however, no model is selected. Rather, the model with the lowest BIC serves as the new constrained model, now with $b + 1$ items relaxed relative to the constrained model in the previous step. The new candidate pool therefore consists of the most constrained model, with $b + 1$ relaxations, and new competitive models with $b + 2$ relaxations. The new competitive models are not only required to have $b + 2$ relaxed items, but the specific item(s) that are relaxed in these models must reflect the relaxed item(s) in the constrained model. In this way, candidate pools in the stepwise approach contain a limited number of models up for consideration at any given time. The new candidate pool is assessed, and if the model with the lowest BIC corresponds to the most constrained model, it is selected. Otherwise, the stepwise process continues until the model with the lowest BIC corresponds to the most constrained model, at which point it is selected. In short, the stepwise method proceeds as follows:

1. The candidate pool is comprised of the most constrained model, with the least number of items relaxed (*b* = 0), as well as competitive models with *b* + 1 items relaxed.
2. The BIC is recorded for each model in the candidate pool.
3. A model is selected from the candidate pool through one of two ways, depending on the model that corresponds to the lowest BIC.
    a. If the model with the lowest BIC corresponds to the most constrained model in the candidate pool, this model is selected.
    b. Otherwise, the model with the lowest BIC is considered the new constrained model, but no model is selected. The process repeats from step 1, but the new constrained model now has *b* + 1 items relaxed, and the new competitive models have *b* + 2 items relaxed.

Alternative Criterions and Stepwise Approaches

While the BIC and its role in isolating non-invariance is a major focus of this experiment, it is not the intention of this paper to imply there are no alternative criterions or tools regularly implemented for model selection procedures. Particular to IRT, the DIF function is a model selection function available through Chalmers' *mirt* package that proceeds in a stepwise manner using Wald and likelihood-ratio tests to isolate non-invariance when a set of predetermined anchor items, constrained to be equal across groups, is specified (Chalmers, 2012). Further, a different information criterion that preceded the BIC developed by Schwarz (1978) is the Akaike Information Criterion, or AIC, developed by Akaike (1974). Unlike the BIC which requires the true model to exist in the candidate pool and has a penalty function very sensitive to sample size

(N), the entropy-based AIC operates best in instances when the true model is not found in the candidate pool, as it tends to instead select the model which most minimizes the mean squared error of estimation (Vrieze, 2012). A host of other model fit indices are used by researchers regularly based on the assumptions that best satisfy or justify the use of a given index. Further, although the stepwise procedure administered in this particular experiment always began from the most constrained model and sequentially relaxed items in an iterative manner, the procedure could have just as well been reversed to begin with a fully relaxed model, sequentially constraining item parameters in an iterative manner. Thus, while the stepwise approach used in this experiment has not been tested before and therefore has limited underlying theory, it is by no means the case that stepwise procedures in general, or other fit indices, have never been explored or implemented.

Simulation and Data Settings

Outcome measures for the experiment were assessed over several scenarios, wherein sample sizes (N), item counts, and magnitude and concentration of DIF varied (see Table 1).

Table 1

Summary of Study Variables

Variable	Level
Total Items	6 or 20
Concentration of DIF Items	15% or 16.7% or 33%
N per group	250 focal and 250 reference 1000 focal and 1000 reference 4000 focal and 4000 reference
DIF Magnitude Settings	*Null:* No difference in parameters *A:* α parameter .5 lower in focal group, no δ parameter DIF *B:* δ parameter .4 higher in focal group, no α parameter DIF *C:* δ parameter .8 higher in focal group, no α parameter DIF *D:* α parameter .5 lower in focal group, δ parameter .4 higher in focal group *E:* α parameter .5 lower in focal group, δ parameter .8 higher in focal group

Note. 84 total conditions x 100 replications = 8,400 samples. DIF = differential item functioning.

In total, 84 simulations were run, and where applicable, the performance of the traditional and stepwise approaches were compared with one another.

Data generated in the experiment was motivated by previously published simulation studies involving information criteria and model selection, as well as theory concerning 2PL models in the context of Item Response Theory. As opposed to studies like Meade and Wright (2012) and Kim et al. (2019), which ran analyses on both simulated and actual data samples, this experiment was run exclusively on simulated data. One objective of the study was to explore if the stepwise approach under consideration hinted in some ways at having a consistency property similar to that of the traditional approach, for applicable magnitude settings. It warrants mentioning, however, that as far as the consistency property is concerned, settings A, B, and C are not necessarily supported by the traditional BIC's consistency property, as the DIF in these settings is expressed through two parameters at a time, as opposed to one. Thus, settings A, B, and C represented unexplored scenarios in relation to the BIC's consistency property. Nevertheless, the experiment was primarily theoretically motivated, and a critical design element of the study was to maintain control over inputs and outputs in order to study the causality of how particular conditions and settings impact the output (Kim, Kang, Choi, & Kim, 2017). Thus, the methodology described herein was explored in strictly simulated settings to maintain such controls.

Sample Size

Each method was tested across three different sample sizes (N): 500, 2,000, and 8,000. N of 500 was selected as the smallest sample size as it is a common value for invariance studies (Meade & Wright, 2012). As is common in simulation studies (Vrieze, 2012; Kim et al., 2019), the sample of 500 was increased in multiple increments, and both methods were tested on sample sizes of 500, 2,000, and 8,000. Borrowing from Meade and Wright (2012), the total sample size across the reference and focal group was evenly split in each simulation.

Item Count

One of the major complications of using traditional BIC-guided model selection in non-invariance investigations is that as the number of items (i) in the test instrument increases, so too does the number of candidate models that must be assessed, and this increase occurs exponentially. The relation between number of items (i) and number of candidate models for two groups, for example, is 2^i. The traditional BIC method is therefore only feasible for non-invariance investigations when the test instrument contains a feasibly small number of items, and consequently, a feasibly small number of candidate models to assess. Two different item counts were used in the experiment. One of the instruments contained six items ($i = 6$), and for this instrument it was therefore possible to evaluate success rates for the traditional approach against the stepwise approach. While many standardized achievement exams often attempt to assess test-takers in respective content domains using 40-50 items (Popham, 1999), it is not uncommon to find subscales comprising high stakes tests with fewer than ten questions (National Center for Education Statistics, 2013). The second instrument phase entailed running the stepwise method on an instrument with a larger item count, $i = 20$. The motivation behind using 20 items, aside from the value being a common item count for invariance investigations (Meade & Wright, 2012), was that 20 items corresponded to 2^{20}, or 1,048,576 candidate models, which was an unfeasibly high number of models for a DIF investigation via the traditional method, but a feasible quantity to explore via the stepwise method, since the stepwise method isolates the true model not by assessing all combinations of non-invariance but by moving through iterative combinations, or steps, of non-invariance until the supposed true model is isolated. Testing the stepwise method on a substantially larger item count was necessary in order to assess the utility of the approach for more practical, real-world test settings, which, on average involve item counts ranging between 40-50 items (Popham, 1999). The experiment was not run on a 40-50 item instrument, however, due to computational and timing constraints that such a large instrument would invoke. Considering that the 20-item instrument still entailed over a million combinations of non-invariance, though, the test was deemed appropriate and telling of how the novel stepwise approach might operate in more realistic, standardized exam settings.

DIF Concentration

The study was broken into three levels of DIF concentration, or items expressing DIF, across the 6-item instrument and the 20-item instrument. Per Table 2, assessing the success rates of both the traditional approach and the stepwise approach on the 6-item instruments across different DIF concentrations (16.7% of items expressing DIF versus 33% of items expressing DIF) allowed for more comprehensive understanding of how the methods worked in varying DIF settings. Finally, as for which specific items expressed DIF in each simulation, this was randomly but identically generated for each instrument by using the same seed for the data generating portions for each simulation in the R software environment.

A noteworthy comment should be made regarding the 16.7% DIF concentration, where DIF was expressed through only one item. An important theoretical feature of the experiment was that in the 16.7% DIF concentration, the theory underpinning the traditional BIC approach could also be applied to the stepwise procedure. While theory on the stepwise approach is, overall, quite limited, a few hypotheses and expectations for both approaches' behavior in detecting the true model can be made in this case.

Table 2

Summary of DIF Variables

Instrument	Proportion of DIF Items	DIF Items	Traditional Approach	Stepwise Approach
6-Item Instrument	16.7%	Item No. 1	Yes	Yes
6-Item Instrument	33%	Item No. 1, 4	Yes	Yes
20-Item Instrument	15%	Item No. 1, 4, 7	No	Yes

Note. DIF = differential item functioning.

When only one item in the true model expresses DIF, it is guaranteed that the true model is among the first batch of candidate models reviewed by the stepwise approach, and if this is the case, regardless of the approach, both procedures are expected to consider the model with the lowest BIC the true model. And as such, this should be the same model in both cases.

Magnitude and Nature of DIF

The magnitude of DIF in each simulation was motivated by that of similar studies (Meade & Wright, 2012) in which chosen items' discrimination parameters and/or difficulty parameters expressed DIF across the sample's focal and reference groups. Per the study design implemented by Meade and Wright (2012), there was a 0.5 difference between groups' discrimination parameters (α), and differences in difficulty parameters (δ) across groups were either small or large, 0.4 or 0.8, respectively. Both the data exhibiting uniform DIF (settings B and C) and nonuniform DIF (settings A, D, and E) was generated in the data generation phase of the experiment, but parameter values varied setting-to-setting. It warrants mentioning, however, that the model selection procedures used in the experiment were engineered to isolate non-invariance that manifests, in the very least, through the discrimination parameter (α). That is to say the procedures used were meant to isolate nonuniform DIF, even though data generated across settings involved both uniform and nonuniform DIF. Regardless, for a test to be considered valid, that is, that the test truly measures the condition it claims to measure, it is necessary to explore and account for DIF (Goetz, et al., 2016). Reinforced by Goetz et al. (2016), however, DIF can exist both at substantial levels that interfere with the scale's ability to measure the construct, and negligible levels that allow the scale to effectively capture the construct. In addition to the five magnitude settings used by Meade and Wright (2012), an additional null setting was tested in which the true model expressed no DIF. The purpose of the null setting was to serve as a crosscheck of sorts to confirm the algorithm used in each simulation could identify the true model before any DIF was introduced. If the algorithm for a given simulation was unable to identify the true model with no DIF present, it would not have been reasonable to assume the algorithm could have isolated the true model in settings that contained DIF. In total, six different magnitude settings were tested for each sample size, for each applicable method: settings A-E, where the magnitude of DIF varied across items' alpha and/or discrimination parameters, and a null setting, where no DIF was present (see Table 1).

Group Mean Differences

Borrowing from Kim et al. (2019), the latent distribution for all simulations in the current study was normally distributed with group mean of 0 for the reference group and group mean of –1 for the focal group. Contrary to Kim et al. (2019) and Meade and Wright (2012), however, a standard deviation of 1 was not used for both groups, but rather only the reference group. The focal group was assigned a standard deviation of 1.2. Thus, the reference group's latent distribution of ~N (0,1) differed in both mean and standard deviation from the focal group's latent distribution of ~N (1,1.2). Different standard deviations were assigned because in real world test settings, it's reasonable to assume different groups of participants would show not just different mean scores, but also different variances in scores.

Simulation and Supplementary Details

Each simulation entailed 100 replications (*z*), and analyses were carried out in R 3.6.2 (R Core Team, 2020), using the *Multidimensional Item Response Theory* (*mirt*) package (Chalmers, 2012) and the *data.table* package (Dowle & Srinivasan, 2019). Documentation pertaining to GDPR and ethics approvals are available in Appendix I, and full reproducible code is available in Appendix II. All models were estimated with the maximum likelihood estimator and 2000 iterations were allotted for each model to converge. Models that did not converge within 2000 iterations were recorded as non-converged by both algorithms.

Outcome Measures

Corrections for non-Convergence

Studies (Vrieze, 2012) have demonstrated that the consistency property of the BIC, its tendency to select the true model with probability approaching 1 as sample size increases, holds true, when a fixed number of candidate models are being considered. Given the criterion's penalty function and sensitivity to sample size, convergence issues were anticipated for simulations run on relatively smaller sample sizes of N=500. Both algorithms allocated 2000 iterations per candidate model for successful convergence. To avoid premature termination in the event of non-convergence for a given model within the allotted 2000 iterations, each algorithm recorded non-converged models.

Model Evaluation

As non-convergence was anticipated, it was necessary to consider results of the traditional and stepwise approaches in context of one another, to avoid reporting biased findings. That is, for the 6-item instrument, of the 100 replications (*z*) conducted via both the traditional approach and the stepwise approach, only replications in which a model converged for both approaches were considered. If, for example, no model converged for the 4th replication ($z = 4$)

of the traditional approach run for the 6-item instrument with DIF concentration 33%, magnitude setting A, and N=500, then the 4th replication (z = 4) for the corresponding stepwise simulation was excluded from final results. This exclusion helped to reduce noise and facilitate more unbiased comparisons between approaches. The true model detection rate for a given simulation, for a given approach, was therefore calculated by:

$$\frac{T_A}{Z_{C.BOTH}}\ , \qquad (3)$$

where, T_A is the number of times the true model (T) was detected in a given approach (A), and $Z_{C.BOTH}$ is the number of replications (Z) that converged for both approaches ($C.BOTH$) for a given simulation. For the 20-item instrument, as only the stepwise approach was implemented, the true model detection rate for a given simulation was calculated based on the number of times the stepwise approach selected the true model, divided by the number of converged replications:

$$\frac{T}{Z_C}\ , \qquad (4)$$

where, T is the number of times the true model was detected in the stepwise approach and Z_C is the number of replications (Z) that converged (C) for that simulation.

Results

This section reports summaries of the true model detection rates for the three different concentrations of DIF tested across two instruments: the 6-item instrument expressing 17% DIF and 33% DIF, and the 20-item instrument expressing 15% DIF. For each instrument, results pertaining to DIF magnitude settings, sample size, and non-convergence instances are presented. Where applicable, results of different instruments are compared against one another.

Results for the 6-Item Instrument, 17% DIF

Table 3

True Model Detection Rate for 6-Item Instrument, 17% DIF

	Sample Size					
	N=500		N=2000		N=8000	
	Approach					
Setting	Traditional	Stepwise	Traditional	Stepwise	Traditional	Stepwise
Null	0.979 (0.015)	0.979 (0.015)	0.990 (0.009)	0.990 (0.009)	1.000 (0.000)	1.000 (0.000)
A	0.076 (0.028)	0.076 (0.028)	0.470 (0.049)	0.470 (0.049)	1.000 (0.000)	1.000 (0.000)
B	0.115 (0.033)	0.115 (0.033)	0.480 (0.049)	0.480 (0.049)	1.000 (0.000)	1.000 (0.000)
C	0.510 (0.051)	0.510 (0.051)	0.990 (0.009)	0.990 (0.009)	1.000 (0.000)	1.000 (0.000)
D	0.010 (0.010)	0.010 (0.010)	0.140 (0.035)	0.140 (0.035)	0.900 (0.030)	0.900 (0.030)
E	0.093 (0.030)	0.093 (0.030)	0.620 (0.049)	0.620 (0.049)	1.000 (0.000)	1.000 (0.000)

Note: Percentages represent number of times the true model was detected by individual approach, divided by the number of mutual converged replications across both approaches. DIF = differential item functioning. DIF expressed through Item No. 1 for settings A-E.

True model detection rates for both approaches implemented on the 6-item instrument expressing 17% DIF are reported in Table 3. Both the traditional and stepwise approaches were implemented for the 6-Item Instrument with 17% DIF expressed through item no. 1. Although true detection rates differed for different sample sizes and magnitude settings, after correcting for non-convergent replications, the two approaches isolated the true model with identical accuracy across all 18 conditions run for this instrument. Standard errors for binomial proportions (Brown, Cai, & DasGupta, 2001) were computed for results for the 6-item instrument with 17% DIF, the 6-item instrument with 33% DIF, and the 20-item instrument with 15% DIF.

Magnitude of DIF

Per Table 3, it is possible that the magnitude of DIF in settings A-E may have a relation with the approaches' success in isolating the true model. With sample size held constant, each approach had the lowest accuracy rates for setting D, where the reference group's α parameter was .5 higher, and its δ parameter .4 lower, than that of the focal group. Both approaches isolated the true model 1.04% of the time in setting D, after correcting for non-converging replications.

Aside from the null setting with zero DIF, the highest accuracy rates for both approaches occurred in setting C, where the δ parameter of the focal group was 0.8 higher than that of the reference group. It was noted that for the N=8000 simulations, the accuracy rates displayed less of a hierarchy across different magnitude settings. That is, when N=8000, the true model was detected 100% of the time in all settings, except in setting D, where the detection rate was 90%. At least as far as the traditional BIC approach's resilience to fluctuating magnitudes is concerned for N=8000, this behavior is likely highlighting the BIC's penalty function described by Equation 2. Assuming its necessary conditions are satisfied, the BIC tends to select the true model with probability 1 as sample size increase, which appears to be the case here.

Sample Size

The traditional BIC approach tended to select the true model with probability approaching 1 as the sample size increased, as relevant theory and literature supported, given necessary assumptions were satisfied and the true model was amongst candidate models. As mentioned previously, given the assumptions and tendencies of the traditional BIC procedure in identifying the true model when only one item expresses DIF, the results in Table 3 should, and do, indicate that the stepwise approach selects the true model with the exact frequency as the traditional BIC approach. Thus, in this capacity, the stepwise procedure behaved as theory would suggest it should.

Instances of non-Convergence

For each magnitude setting and sample size, each approach was run over the course of 100 replications. Instances of non-convergence occurred only when running the traditional approach for the N=500 sample size (see Table 4).

Table 4

Instances of Non-Convergence for the 6-Item Instrument, 17% DIF

Sample Size, N=500		
	Approach	
Magnitude Setting	Traditional	Stepwise
Null	7	0
A	8	0
B	4	0
C	4	0
D	4	0
E	4	0

Note: Counts shown represent the number of replications where model did not converge, and were therefore omitted from model accuracy computation rates for both approaches.

As theory pertaining to the BIC mandates a sufficiently large sample size for the criterion to probabilistically select the true model, it is reasonable that the smallest sample size tested, N=500, saw instances of non-convergence. Of note was that the stepwise approach experienced no such instances for the corresponding simulations, but non-converged replications were nonetheless removed from computation of both approaches' accuracy rates, where applicable.

Results for the 6-Item Instrument, 33% DIF

True model detection rates for the traditional and stepwise approaches implemented on the 6-item instrument expressing 33% DIF are reported in Table 5. Both approaches were run separately across three sample sizes and five DIF magnitude settings. As the null settings remained unchanged from those run for the 6-item instrument expressing 17% DIF, null settings were not run again for the 6-item instrument expressing 33% DIF. The true model expressed 33% DIF via items 1 and 4 for settings A-E. Both approaches' results suggested varying true model detection rates for different sample sizes and magnitude settings.

Table 5

True Model Detection Rate for 6-Item Instrument, 33% DIF

	Sample Size					
	N=500		N=2000		N=8000	
	Approach					
Setting	Traditional	Stepwise	Traditional	Stepwise	Traditional	Stepwise
Null	0.979 (0.015)	0.979 (0.015)	0.990 (0.009)	0.990 (0.009)	1.000 (0.000)	1.000 (0.000)
A	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.290 (0.045)	0.290 (0.045)
B	0.000 (0.000)	0.000 (0.000)	0.540 (0.049)	0.460 (0.049)	1.000 (0.000)	1.000 (0.000)
C	0.604 (0.049)	0.510 (0.049)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
D	0.000 (0.000)	0.000 (0.000)	0.190 (0.039)	0.190 (0.039)	0.880 (0.032)	0.880 (0.032)
E	0.232 (0.045)	0.202 (0.041)	0.900 (0.030)	0.900 (0.030)	1.000 (0.000)	1.000 (0.000)

Note: Percentages represent number of times the true model was detected by individual approach, divided by the number of mutual converged replications across both approaches. DIF = differential item functioning. DIF expressed through Item No. 1 and Item No. 4 for settings A-E.

Unlike the previous 6-Item Instrument, where the stepwise approach performed with identical accuracy to the traditional approach in all conditions, for the instrument with 33% DIF, this matching of performance between approaches only occurred for specific settings and sample sizes. Particularly, the approaches began showing identical detection rates only for the relatively larger sample sizes of N=2000 and N=8000. Another interesting observance was that although the stepwise approach overall operated more efficiently, computationally speaking, it was the traditional approach that outperformed the stepwise approach when the two methods did not select the true model at identical rates.

Magnitude of DIF

Per Table 5, the traditional and stepwise approach's true model detection rates varied across the different magnitude settings used for the 6-Item Instrument with 33% DIF. Sample size held constant, each approach had the lowest accuracy rates in setting A, where the reference group's α parameter was .5 higher than that of the focal group. In the smaller sample sizes of N=500 and N=2000, neither approach succeeded in isolating the true model in setting A. For the sample size of N=8000, the true model detection rate for both approaches was 29%. As it pertains to the traditional BIC method, it was interesting to note that when the concentration of DIF was increased from 17% across one item to 33% across two items, the BIC's penalty function did not correct for ML estimation as successfully as it did for the 17% DIF instrument. Further investigation may be warranted for exploring if the BIC is sensitive to certain item parameters more than others when the concentration of non-invariance varies in IRT non-invariance investigations. As for the highest accuracy rates observed for this instrument, overall, both approaches selected the true model with highest frequency in setting C, where the δ parameter of the focal group was 0.8 higher than that of the reference group. As far as overall fluctuations in magnitude were concerned, it was noted that for the N=500 and N=8000 simulations, the accuracy rates displayed less of a hierarchy across different settings. That is, when N=500, neither approach detected the true model for settings A, B, and D (detection rates of 0%). When N=8000, both approaches had 100% detection rates for settings B, C, and E.

Sample Size

In the 6-Item Instrument with 33% DIF, the traditional BIC approach tended to select the true model with probability approaching 1 as sample size increased, as relevant theory and literature suggested it should, so long as conditions were met and the true model was in the candidate pool. As sample size increased, the stepwise approach also tended to select the true model with greater accuracy. Further, the stepwise approach displayed identical detection rates as the traditional approach for almost all settings, but only for the relatively larger sample sizes (N=2000 and N=8000). In setting B of the N=2000 simulation, where the δ parameter of the focal group was 0.4 higher than that of the reference group, not only did the two approaches select the true model at different frequencies, but the traditional approach outperformed the stepwise approach. As no theory exists for the model selection properties of the stepwise approach in this capacity , it is not clear why or how the approach was not as successful in selecting the true model in not only setting B of the N=2000 sample size, but also settings C and E of the N=500 sample size.

Instances of non-Convergence

100 replications were run for the 6-Item Instrument expressing 33% DIF. Just as the 6-Item Instrument with 17% DIF, instances of non-convergence occurred only for the traditional approach and only for the N=500 sample size (see Table 6). Unlike the previous 6-Item Instrument, where the stepwise approach performed not only faster and more efficiently, but with identical accuracy as the traditional method, per Table 6, this was not the case for the 6-Item Instrument with 33% DIF. Although the stepwise approach experienced no issues with non-convergence, it did not select the true model with the same accuracy as the traditional approach in a handful of settings. While the two approaches' accuracy rates are within a 10% margin of one another, it is not clear why the stepwise approach did not select the true model as often as the traditional method, or what methodological adjustments may be necessary.

Table 6

Instances of Non-Convergence for the 6-Item Instrument, 33% DIF

Sample Size, N=500		
	Approach	
Setting	Traditional	Stepwise
Null	7	0
A	3	0
B	1	0
C	4	0
D	5	0
E	1	0

Note: For each set of 100 replications, the counts shown represent the number of replications where a model did not converge. These replications were omitted from model accuracy computation rates for both approaches.

Results for the 20-Item Instrument, 15% DIF

Table 7

True Model Detection Rate for 20-Item Instrument, 15% DIF

	Stepwise Approach		
	Sample Size		
Setting	N=500	N=2000	N=8000
Null	0.970 (0.017)	0.980 (0.014)	1.000 (0.000)
A	0.000 (0.000)	0.060 (0.023)	0.940 (0.023)
B	0.000 (0.000)	0.270 (0.044)	1.000 (0.000)
C	0.450 (0.049)	1.000 (0.000)	1.000 (0.000)
D	0.010 (0.009)	0.680 (0.046)	1.000 (0.000)
E	0.190 (0.039)	0.970 (0.017)	1.000 (0.000)

Note: Percentages represent times the true model was detected by the stepwise approach, divided by 100 converged replications. DIF = differential item functioning. DIF expressed through Item No. 1, Item No. 4, and Item No. 7 for settings A-E.

True model detection rates for the stepwise approach, implemented on the 20-item instrument are reported in Table 7. The traditional approach was not tested due to the computational and timing constraints. The true model in the null settings expressed no DIF, whereas the true model in remaining settings, A-E, expressed 15% DIF through items 1, 4, and 7.

Magnitude of DIF

Sample size held constant for the 20-item instrument, the stepwise approach had the lowest accuracy rates for setting A, where the reference group's α parameter was .5 higher than that of the focal group. Again, it is unclear if the stepwise BIC approach has a particular sensitivity, or lack thereof, to certain item parameters, namely the alpha parameter, in context of 2PL models. As the traditional approach was not administered for this instrument, it is not possible to compare the stepwise approach's sensitivity to magnitude settings to that of the traditional approach, as was done for the 6-item instrument. Excluding the null settings, the setting that saw the highest true model detection rate across all sample sizes was setting C, where the reference group's δ parameter was .8 lower than that of the focal group. As was the case for the stepwise approach run on the 6-item instruments, it was noted that for the largest sample size of N=8000, the accuracy rates displayed less of a hierarchy across different magnitudes settings. That is, when N=8000, the stepwise approach detected the true model with 100% accuracy for all settings except for setting A, where it was detected with 94% accuracy.

Sample Size

Across all six settings, there appeared to exist an overall positive relationship between the true model detection rate and sample size for the stepwise approach used on the 20-item instrument. As the traditional approach was not used for the 20-item instrument, no supporting theory or literature was available to supplement the stepwise approach's behavior in context of that of the traditional method. The only sample size for which a true model was not detected to any extent was the sample size of N=500, where the detection rates for settings A and B were both 0%. Whereas true model detection rates varied for settings in the N=500 and N=2000 simulations, however, the true model was detected with 100% accuracy for all but one setting in the N=8000 simulations. Further research is required to determine if the stepwise approach has a penalty function similar or identical to that of the traditional BIC approach, where probability of isolating the true model increases as sample size does.

Instances of non-Convergence

There were no instances of non-convergence for the stepwise method used on the 20-item instrument. Thus, all true model detection rates shown in Table 7 represent the number of times the true model was correctly selected out of 100 converged replications. Further research might be warranted to determine why even in relatively smaller sample sizes like N=500, the stepwise approach does not experience convergence issues. Additionally, of interest would be to determine what sample size, if any, prevents the stepwise procedure from converging.

Discussion

This section focuses on how, based on the results of this experiment, the stepwise procedure may be of use in non-invariance investigations for large-scale testing agencies. Next steps for follow-up investigations, and limitations of the current experiment, are also discussed. Finally, research questions of this paper are revisited and concluding thoughts are offered.

Consequences for Large-scale Testing

For the instruments in which the stepwise procedure's performance could be directly compared with that of the traditional procedure, the findings of this experiment were that at a sufficiently large sample size, N=8000 in this case, the stepwise procedure performed with the exact accuracy as the traditional procedure, in a fraction of the time. While the stepwise procedure sometimes had slightly lower accuracy rates than the traditional approach for smaller sample sizes of N=500 and N=2000, typical large-scale tests would rarely present insufficient sample size as an issue. The Graduate Record Examination (GRE), for example, was administered to more than 500,000 test takers in the 2016-2017 calendar year, alone (Educational Testing Service, 2017). For a test like the GRE, which adheres to a 2PL model and undoubtedly hosts several thousand examinees per implementation, a stepwise procedure for isolating DIF could be administered without running the risk of yielding less accurate results than a traditional BIC procedure. It should be reiterated, however, that it is very unlikely in the first place that the traditional procedure would be used by agencies like the Educational Testing Service to isolate non-invariance in sections of the GRE. The computer-delivered GRE has 20-item sections for both the Quantitative Reasoning and Verbal Reasoning portions of the exam (Educational

Testing Service, 2020). Implementing the traditional BIC approach for a reference and focal group for either of these sections would require the procedure to work through 2^{20} combinations of invariance, demanding unfeasible amounts of manpower, time, and resources. Should the Educational Testing Service seek to investigate non-invariance using the BIC, however, the stepwise approach explored in this experiment could presumably save the agency a considerable amount of time and resources.

Additionally, the potential utility of the stepwise approach for instruments where DIF is expressed through a single item should not be understated. In the 6-item instrument expressing 17% DIF in this experiment, it was observed that the stepwise procedure performed identically to the traditional approach in all conditions and cases. Although other stepwise procedures, such as Chalmers' DIF argument, available through the *mirt* package (Chalmers, 2012), have not been fully investigated in academic publications as of yet, the results of this experiment suggest that the stepwise BIC procedure operates more robustly compared to the DIF function. Further research and testing of this paper's BIC procedure against Chalmers' (2012) DIF function is of course required, however.

Further, although the stepwise procedure appears to operate with identical accuracy as the traditional approach only in larger sample sizes, that should not dismiss the potential utility of the stepwise procedure for investigations involving reasonably smaller sample sizes. As demonstrated in this paper and supporting literature, the traditional BIC's penalty criterion is clearly sensitive to sample size and selects the true model with greater probability as sample size increases. Sample size is central to the BIC's theory, and if a given sample size is drastically insufficient, it is of course more reasonable to conduct a non-invariance investigation with different criterions or procedures altogether, like the Akaike Information Criterion (AIC) or the *mirt* package's DIF function (Chalmers, 2012). However, based on the preliminary results of this experiment, under specific conditions the stepwise procedure may still prove useful for a test implemented on a smaller sample if the test agency sought to sacrifice a predetermined level of accuracy for a computationally more efficient model selection procedure. For example, when the stepwise procedure was run on the N=2000 sample size for both 6-item instruments, it selected the true model with underwhelming accuracy in some conditions, but higher than 90% accuracy for other conditions. Contingent on further research, it's possible that a sample size between N=2000 and N=8000 could be pinpointed, for example, where accuracy rates were secured within an acceptable range, albeit lower than 100%. Many agencies that administer high stakes tests already publish and distribute technical reports pertaining to the test's measurement error (Newton, 2005). If the stepwise BIC procedure was deemed appropriate, for a sufficient but relatively smaller sample size, the agency in question could just as easily account for possible deviations in accuracy rates through these technical reports.

Regardless of whether the sample size is beyond sufficient or barely sufficient for the stepwise procedure to yield worthy results, it should be reemphasized that model selection is a comprehensive process that involves many different elements of evaluation and assessment of a model (Brown T. A., 2015). That is, it is not the objective of this paper to imply that a scale's non-invariance can be isolated through a single criterion, like the BIC. Even if the stepwise procedure were somehow developed to operate with 100% accuracy in isolating the true model in all conditions and settings, the fact remains that far more goes into selecting a model, or in this case isolating non-invariance, than solely the fit of a model, which is what the BIC reveals. Many test agencies, namely the GRE, already implement a host of procedures and tools for ensuring fair tests across different groups, including but not limited to DIF analysis (Educational

Testing Service, 2017). The aim of this paper is to suggest that, contingent on further research and understanding of its assumptions, the stepwise BIC procedure may prove a useful and worthwhile addition to the non-invariance toolbox that large-scale agencies already use in the test design and analysis process.

Next Steps and Limitations of the Current Experiment

Although based on the preliminary results of this experiment, the stepwise procedure shows worthwhile potential for use in future non-invariance investigations, there are limitations in the current experiment that must be addressed prior to any such implementations. First and foremost, the current study tested the stepwise procedure in an exclusively simulated environment and testing the procedure on actual test data is necessary to verify that the stepwise procedure performs identically to the traditional procedure for a sufficiently large sample size. Additionally, although the traditional procedure was not tested on the 20-item instrument due to computational and timing constraints of the current study, in order to verify that the stepwise procedure yields identical results as the traditional approach, regardless of number of observations, or items (*i*), the traditional approach must be tested on the 20-item instrument for comparison purposes. Further, as evident from the results of the 6-item instrument when DIF concentration was increased from 17% to 33%, there was an apparent change in accuracy rates for both the traditional and stepwise procedures, and it is for instruments in which DIF is expressed through more than one item where theory pertaining to the stepwise BIC procedure is lacking. Further investigations should consider testing the accuracy of the approaches over varying concentrations of DIF, as opposed to only two concentrations, as was done in the current study. That is to say results were not conclusive for how or why the accuracy rates of the approaches shift over varying concentrations of DIF, or if there is a relationship between accuracy and DIF concentration. Last but not least, an obvious but important limitation of the current study was that no theory or literature currently pertains to how the stepwise BIC procedure is expected to behave as sample size increases when more than one item expresses DIF. Thus, while recording results for the procedure, it was not possible to say conclusively whether the stepwise procedure was behaving according to its theoretical assumptions. Therefore, further investigation and research is warranted to determine if the stepwise procedure has similar assumptions to the traditional BIC approach, if it has any assumptions at all.

Revisiting Objectives of the Current Experiment

This experiment sought, among other objectives, to explore whether it was possible to isolate the source of non-invariance in a candidate model pool using a stepwise approach built around the consistency property of the traditional BIC. While further investigation is warranted in many areas, this experiment found that, at least for a 6-item instrument with different concentrations and magnitudes of DIF, the stepwise approach did isolate the source of non-invariance with the exact accuracy levels as the traditional BIC approach. Further, results from the experiment showed that higher accuracy rates were obtained for the 6-item instrument when the concentration of DIF was isolated to one item (17%) versus two items (33%), as well as when the magnitude of DIF was expressed through a larger difficulty parameter ($\Delta\delta = 0.8$). In spite of this, it is not possible to say conclusively if or how magnitude and concentration of DIF impact either approach's ability to detect the true model. Further investigation is warranted in this respect. It was established, however, that all things constant, N=8000 served as a sufficiently large sample size to facilitate identical accuracy rates between the stepwise and traditional approach for the 6-item instrument. These results motivate the need for follow-up investigation and testing across additional sample sizes, test responses, and settings and concentrations of DIF to further build out the utility of the stepwise procedure. Contingent on additional studies, the results presented herein suggest that the stepwise procedure stands to benefit efforts of high stakes test agencies in their endeavors to craft and analyze fairer examinations.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*, 716 –723. doi:10.1109/TAC.1974.1100705

Ayala, R. d. (2009). *The Theory and Practice of Item Response Theory.* New York City: The Guilford Press.

Brown, C., Templin, J., & Cohen, A. (2015). Comparing the Two- and Three-Parameter Logistic Models via Likelihood Ratio Tests: A Commonly Misunderstood Problem. *Applied psychological measurement, 39*(5), 335–348. doi:10.1177/0146621614563326

Brown, L. D., Cai, T. T., & DasGupta, A. (2001). Interval Estimation for a Binomial Proportion. *Statistical Science, 16*(2), 101-133. doi:10.1214/ss/1009213286

Brown, T. A. (2015). *Confirmatory Factor Analysis for Applied Research* (2nd ed.). New York City: The Guilford Press.

Chalmers, P. (2012). mirt: A multidimensional item response theory package for the R environment. R Package Version 1.3.1. *Journal of Statistical Software, 48*(6), 1-29. doi:10.18637/jss.v048.i06

Dowle, M., & Srinivasan, A. (2019). data.table: Extension of `data.frame`. R package version 1.12.8. Retrieved from https://CRAN.R-project.org/package=data.table

Educational Testing Service. (2017). *A Snapshot of the Individuals Who Took the GRE General Test, July 2012- June 2017.* ETS.

Educational Testing Service. (2020). *Computer-delivered GRE® General Test Content and Structure*. Retrieved April 13, 2020, from www.ets.org: https://www.ets.org/gre/revised_general/about/content/computer

Froese-Germain, B. (2001). Standardized Testing + High-Stakes Decisions = Educational Inequity. *Interchange, 32*, 111–130. doi:10.1023/A:1011985405392

Goetz, C. G., Liu, Y., Stebbins, G. T., Wang, L., Tilley, B. C., Teresi, J. A., . . . Luo, S. (2016). Gender-, age-, and race/ethnicity-based differential item functioning analysis of the movement disorder society-sponsored revision of the Unified Parkinson's disease rating scale. *Movement disorders : official journal of the Movement Disorder Society, 31*(12), 1865-1873. doi:10.1002/mds.26847

Kelderman, H., & Rijkes, C. P. (1994). Loglinear multidimensional IRT models for polytomously scored items. *Psychometrika, 59*, 149–176. doi:10.1007/BF02295181

Kim, B., Kang, B., Choi, S., & Kim, T. (2017). Data modeling versus simulation modeling in the big data era: case study of a greenhouse control system. *SIMULATION, 93*(7), 579–594. doi:10.1177/0037549717692866

Kim, S.-H., Cohen , A. S., Cho, S.-J., & Cho , H. (2019). Use of Information Criteria in the Study of Group Differences in Trace Lines. *Applied Psychological Measurement, 43*(2), 95-112. doi:10.1177/0146621618772292

Meade, A. W., & Wright, N. A. (2012). Solving the Measurement Invariance Anchor Item Problem in Item Response Theory. *Journal of Applied Psychology, 97*(5), 1016-1031. doi:10.1037/a0027934

Mislevy, R. (2009). *Validity from the perspective of model-based reasoning (CRESST Report no. 752).* Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing. Retrieved from https://files.eric.ed.gov/fulltext/ED507085.pdf

National Center for Education Statistics. (2013). *2013 Mathematics Assessment.* Institute of Education Sciences,. National Assessment of Educational Progress (NAEP).

Newton, P. E. (2005). Threats to the professional understanding of assessment error. *Journal of Education Policy, 20*(4), 457-483. doi:10.1080/02680930500132288

Nguyen, T. H., Han, H.-R., Kim, M. T., & Chan, K. S. (2014). An introduction to item response theory for patient-reported outcome measurement. *The patient, 7*(1), 23–35. doi:10.1007/s40271-013-0041-0

Popham, W. J. (1999). Why Standardized Tests Don't Measure Educational Quality. *Using Standards and Assessments, 56*(6), 8-15.

R Core Team. (2020). R: A Language and Environment for Statistical Computing. Vienna, Austria. Retrieved from https://www.R-project.org

Ricarte, T., Cúri, M., & von Davier, A. (2017). Modeling Accidental Mistakes in Multistage Testing: A Simulation Study. In M. Wiberg, S. Culpepper, R. Janssen, J. González, & D. Molenaar, *Quantitative Psychology: The 82nd Annual Meeting of the Psychometric Society* (pp. 55-66). Springer.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics, 6*(2), 461-464. doi:10.1214/aos/1176344136.

Sclove, S. L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika 52, 333–343, 52*, 333-343. doi:10.1007/BF02294360

Vrieze, S. I. (2012). Model selection and psychological theory: a discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychological methods, 17*(2), 228-243. doi:10.1037/a0027127

Appendix I

GDPR Documentation

NOTIFICATION FORM (ENGLISH TRANSLATION) – NSD Form

- Personal data
- Types of data
- Project information
- Responsibility for data processing
- Sample and criteria
- Third persons
- Documentation
- Other approvals
- Processing
- Information security
- Duration of project
- Additional information

Personal data

Which personal data will be processed?

Personal data are any data about an identified or identifiable natural person (data subject). Pseudonymized data are also considered personal data. “Pseudonymization” means processing collected data in way that the data can no longer be linked to individual persons, without the use of additional information. This usually involves removing identifiable information such as name, national ID number, contact details etc. from the collected data and giving each data subject a code/number. A scrambling key is the file/list of names and codes that makes it possible to identify individuals in the collected data. The scrambling key should be stored separately form the rest of the data. NB: processing pseudonymized data is still considered processing personal data, even if you do not have access to the scrambling key, and even if the scrambling key is being stored by an external party, such as SSB, the National registry etc.

Types of data

Name. First name and surname. No

National ID number or other personal identification number 11-digit personal identifier, D number, or other national identification number. No

Date of birth. N/A

Address or telephone number. No

Email address, IP address or other online identifiers. An email address is a unique address that is assigned to the user of an electronic mail service. An IP address is a unique address that is assigned to a device (e.g. a computer) in a computer network like the Internet. Dynamic IP addresses may also be considered personal data in certain cases. Cookies are an example of an online identifier. NB! If you are going use an online survey, and the service provider (data processor) will have access to email addresses or IP addresses, you must indicate this here. No

Photographs or video recordings of persons. Photographs and video recordings of faces are usually considered to be personal data. No

Audio recordings of persons. Audio recordings where personal data are recorded and/or where there exists a scrambling key that links the audio recordings to individual persons on the recordings. The voice of the person speaking may be considered personal data in combination with other background information. No

GPS data or other geolocation data. Data which indicate the geographical location of a person. No

Demographic data that can identify a natural person. E.g. a combination of information such as municipality of residence, workplace, position, age, gender etc. No

Genetic data. Personal data relating to the inherited or acquired genetic characteristics of a natural person, which give unique information about the physiology or health of that person. No

Biometric data. E.g. fingerprint, handprint, facial form, retina and iris scan, voice recognition, DNA. No

Other data that can identify a natural person. No

Will special categories of personal data or personal data relating to criminal convictions and offences be processed?

Racial or ethnic origin. This includes belonging to an ethnic group, population, cultural sphere or society that has common characteristics. For example, information that a person is Sami is not considered to say anything about race, but it says something about ethnicity. No

Political opinions. That a person is a member of a political party and/or what a person voted in an election, including political opinions and beliefs. However, this does not include information that a person is a conservative, radical or labor party supporter. No

Religious beliefs. That a person is a member of a religious organization/congregation. This does not include information that a person has a subscription to a religious newspaper. No

Philosophical beliefs. That a person is a member of a philosophical association, or that a person believes that knowledge is acquired through logical speculation and observation. No

Trade Union Membership. That a person is a member of a trade union that organizes employees within the same industry/subject area, e.g. LO, NTL, NAR etc. No

Health data. Personal data concerning a natural person's physical or mental health, including use of healthcare services. No

Sex life or sexual orientation. A person's sexual orientation (homosexual, lesbian, bisexual etc.) and/or sexual behavior (e.g. that a personal has been unfaithful, indecent exposure, offensive gestures/language) No

Criminal convictions and offences. Personal data concerning convictions and offences or related to security measures. No

Project Information

Title

Model Selection using a Stepwise Bayesian Information Approach in Multiple Group Models with Binary Data

Project description

Give a description of the project's scientific purpose/research question

There are many criteria available to a researcher to dictate model selection in Item Response Theory (IRT) non-invariance investigations. The Bayesian Information Criterion (BIC) is of particular interest due to its consistency property, which, given a fixed number of models and met conditions, allows for the true model to be selected with probability approaching 1 as the sample size increases. Current theory supports using the BIC as a possible selection criterion in specific circumstances in which every candidate model is assessed, and consequently requires the number of candidate models to be fixed to a feasible quantity. In non-invariance investigations, however, the amplified increase in candidate models often does not allow for the BIC to assess each and every model, as it is usually no longer feasible. This simulation investigation seeks to study the ability of the BIC to select the true model in several non-invariance simulations across a spectrum of sample sizes, number of items, and varying concentrations of item-level non-invariance. As several trials in the simulation involve an unfeasibly high number of candidate models, a stepwise function will be designed to sift through candidate models to the extent possible, sequentially relaxing item parameters, and ultimately selecting the true model upon isolating a BIC value that no longer improves. As the stepwise approach does not require each and every model to be assessed, it is the objective of this study to gain insight into the accuracy and behavior of this simplified, stepwise BIC function in isolating the true non-invariant model, and in cases permissible, to compare the selection abilities of the stepwise BIC approach to that of its traditional BIC counterpart.

Subject area

- Social sciences

- Technological sciences

Will the collected personal data be used for other purposes, in addition to the purpose of this project? N/A

Personal data should only be processed for specified, explicit and legitimate purposes. This means that each purpose for processing personal data must be identified and described clearly and accurately. In order for a purpose to be considered legitimate, it must also be in accordance with ethical and legal norms.

Explain why it is necessary to process personal data. N/A

Explain why the personal data are adequate, relevant and limited to what is necessary for the purposes for which they are being processed. This includes limiting the amount of collected data to that which is necessary to realize the purposes of data collection. N/A

External funding

- The Research Council of Norway (Norges forskningsråd - NFR) N/A
- Public authorities. E.g. research commissioned by a ministry N/A
- Other. E.g. funding from a pharmaceutical company or from private actors N/A

Type of project

- Research Project and PhD thesis
- [✓ Student project, Master's thesis]
- Student project, Bachelor's thesis
- Other student projects

Responsibility for data processing

Both the student and the supervisor will not handle personal data as data for the project will be simulated.

Data controller N/A

The institution responsible for the processing of personal data. The data controller determines the purposes for which, and the manner in which, personal data are processed.

Project leader (research assistant/ supervisor or research fellow/ PhD candidate) Nasseem Hessami - Master Student, UiO

Björn Andersson – supervisor, Associate Professor, CEMO, UiO, bjorn.andersson@cemo.uio.no

Will the responsibility for processing personal data be shared with other institutions (joint data controllers)? N/A

If two or more institutions together decide the purposes for which personal data are processed, they are joint data controllers.

Joint data controllers N/A

Institution

Institution not found in the list

Institution

Country

Postal address

Email address

Telephone number

Sample and criteria

Whose personal data will be processed?

You must describe each group of people whose personal data you will be processing. Add and describe each sample individually. N/A. Personal data will not be collected, as this is a simulation study.

Sample 1 Describe the sample N/A

Recruitment or selection of the sample N/A

Describe how the sample will be recruited and how initial contact with the sample will be made. For example, whether you will make initial contact during fieldwork or via your own network, or whether a school, hospital or organization will contact its pupils, patients or members on your behalf. If the sample will not be recruited but will be selected from a registry or an administrative system etc., describe how the selection will be carried out and what the selection criteria will be.

Age N/A

Will you include adults (18 y.o. +) who do not have the capacity to consent?

i.e. the person has reduced capacity or lacks capacity to consent. For example, the person may have mental/cognitive impairment, significant physical/emotional ailments, or may be unconscious, conditions which make it difficult or impossible for the person to gain sufficient understanding in order to give valid consent. The central aspect is whether the person is capable of understanding the purpose of the processing/project in question, and of understanding potential positive and negative consequences (immediate and long-term).

Types of personal data - sample 1 N/A. Personal data will not be collected, as this is a simulation study.

Name N/A

National ID number or other personal identification number N/A

Date of birth N/A

Address or telephone number N/A

Email address, IP address or other online identifier N/A

Photographs or video recordings of persons N/A

Audio recordings of persons N/A

GPS data or other geolocation data N/A

Demographic data that can identify a natural person N/A

Genetic data N/A

Biometric data N/A

Other data that can identify a natural person N/A

Methods /data sources - sample 1. N/A. Personal data will not be collected, as this is a simulation study.

Select and/or describe the method(s) for collecting personal data and/or the source(s) of data N/A

Personal interview N/A

Group interview Online survey Paper-based survey N/A

Participant observation - Non-participant observation N/A

Field experiment / field intervention N/A

Web-based experiment N/A

Tests for pedagogical research / psychological tests N/A

Medical examination and/or physical tests N/A

Human biological material N/A

Social media – open forum N/A

Social media – closed forum N/A

Discussion board/forum for online newspapers/online debates N/A

Big data N/A

Medical records N/A

Biobank N/A

Data from another research project N/A

Other N/A

Statistics Norway - SSB N/A

Criminal records (Det sentrale straffe- og politiopplysningsregisteret, SSP) N/A

Medical Birth Registry of Norway (Medisinsk fødselsregister, MFR) N/A

Norwegian Registry of Pregnancy Termination (Register over svangerskapsavbrudd) N/A

Norwegian Cardiovascular Disease Registry (Hjerte- og karregisteret) N/A

Norwegian Cause of Death Registry (Dødsarsaksregisteret, DÅR) N/A

Norwegian Prescription Database - NorPD (Reseptregisteret) N/A

Norwegian Immunisation Registry (Nasjonalt vaksinasjonsregister, SYSVAK) Norwegian Surveillance System for Communicable Diseases (Meldesystem for smittsomme sykdommer, MSIS) N/A

Norwegian Surveillance System for use of antibiotics and healthcare related infections (Norsk overvåkingssystem for antibiotikabruk og helsetjenesteassosierte infeksjoner, NOIS) N/A

Norwegian Surveillance System for Antimicrobial Drug Resistance (Norsk overvåkingssystem for antibiotikaresistens hos mikrober, NORM) Norwegian Surveillance System for Virus Resistance (Norwegian Surveillance System for Virus Resistance, RAVN) N/A Norwegian Patient Registry (Norsk pasientregister, NPR) IPLOS-registeret Kommunalt pasient- og brukerregister (KPR) N/A Cancer registry of Norway (Kreftregisteret) N/A

Genetic Mass Survey of Newborns (Genetisk masseundersøkelse av nyfødte) N/A

Reseptformidleren N/A

Forsvarets helseregister N/A

Helsearkivregisteret N/A

Helseundersøkelsen i Nord Trøndelag (HUNT) N/A

Tromsø-undersøkelsen N/A

SAMINOR N/A

Den norske mor og barn undersøkelsen (MoBa) N/A

Nasjonalt register for langtids mekanisk ventilasjon N/A

Nasjonalt kvalitetsregister for barnekreft N/A

Norsk Kvalitetsregister Øre-Nese-Hals –Tonsilleregisteret N/A

Norsk vaskulittregister biobank (NorVas) N/A

Norsk Parkinsonregister biobank N/A

Norsk karkirurgisk register (NORKAR) N/A

Norsk hjertinfarkregister N/A

Gastronet N/A

Norsk register for analinkontinens N/A

Nasjonalt barnehofteregister N/A

Norsk kvalitetsregister for artrittsykdommer (NorArtritt) N/A

Norsk nakke- og ryggregister N/A

Nasjonalt korsbåndregister N/A

Nasjonalt register for leddproteser N/A

NorKog N/A

Norsk MS-register og biobank N/A

Nasjonalt register for KOLS N/A

Nasjonalt kvalitetsregister for lymfom og lymfoide leukemier N/A

Nasjonalt kvalitetsregister for lungekreft N/A

Nasjonalt kvalitetsregister for føflekkreft N/A

Nasjonalt kvalitetsregister for brystkreft N/A

Nasjonalt kvalitetsregister for prostatakreft N/A

Nasjonalt kvalitetsregister for tykk- og endetarmskreft N/A

Nasjonalt register for ablasjonsbehandling og elektrofysiologi i Norge (ABLA NOR) N/A

Norsk register for invasiv kardiologi (NORIC) N/A

Norsk hjertesviktregister N/A

Norsk pacemaker- og ICD- register N/A

Nasjonalt kvalitetsregister for gynekologisk kreft N/A

Norsk register for gastrokirurgi (NoRGast) N/A

Nasjonalt kvalitetsregister for behandling av spiseforstyrrelser (NorSpis) N/A

Information - sample 1

Will you inform the sample about processing their personal data? N/A

How? N/A

Written information (on paper or electronically)

Oral information

See what you must give inform about and preferably use our template for the information letter.

Information should be given in writing or electronically. Only in special cases is it applicable to give oral information, if a participant asks for this. See what you must give information about.

Upload information letter N/A

Upload copy of oral information N/A

Explain why the sample will not be informed about the processing of their personal data. N/A. No personal data will be collected or processed. + Add sample

Third persons

Will you be processing personal data about third persons? This includes data about persons who are not included in the sample/are not participating in the project; information provided by a data subject that relates to another identified or identifiable natural person. Examples of this are when a data subject is asked about their mother's and father's education or country of origin, or when pupils are asked about their teacher's teaching methods. N/A. Personal data will not be collected, as this is a simulation study.

Describe the third persons N/A

Types of personal data about third persons N/A

Name N/A

National ID number or other personal identification number N/A

Date of birth N/A

Address or telephone number N/A

Email address, IP address or other online identifiers N/A

Photographs or video recordings of persons N/A

Demographic data that can identify a natural person N/A

Genetic data N/A

Biometric data N/A

Other data that can identify a natural person N/A

Which sample will provide information about third persons? N/A

Will third persons consent to the processing of their personal data? N/A

Will third persons receive information about the processing of their personal data? N/A

Explain why third persons will not be informed. N/A

Documentation

Total number of data subjects in the project (Data subjects: persons whose personal data you will be processing)

- [✓ N/A]
- 1-99
- 100-999
- 1000-4999
- 5000-9999
- 10.000-49.999
- 50.000-100.000
- 100.000+

How can data subjects get access to their personal data or how they can have their personal data corrected or deleted? N/A

Rights of data subjects (participants) include the right to access one's own personal data and to receive a copy of one's data if asked for. A data subject can request that their personal data are corrected if they feel that the information is wrong or lacking, and the data subject can withdraw consent and request that their personal data are deleted. Give a short description of the procedure for how a data subject can get access to their personal data, and how they can have their personal data corrected or deleted.

Other approvals

Will you obtain any of the following approvals or permits for the project? The data for the project will be simulated. Therefore, no known approval or permits are required as no personal data will be collected and/or used.

Indicate if you will obtain any of the following approvals or permits in order carry out the project.

- Ethical approval from The Regional Committees for Medical and Health Research Ethics (REC).
- Confidentiality permit (exemption from the duty of confidentiality) from the Regional Committees for Medical and Health Research Ethics (REC)
- Approval from own management for internal quality-assurance and evaluation of health services (intern kvalitetssikring) (The Health Personnel Act § 26)
- Confidentiality permit (exemption from the duty of confidentiality) from the Norwegian Directorate of Health, for quality-assurance and evaluation of health services (kvalitetssikring) (The Health Personnel Act § 29b)
- Biobank
- Confidentiality permit (exemption from the duty of confidentiality) from Statistics Norway (SSB). Statistics Norway has the authority to grant a confidentiality permit for the data that they manage, e.g. data about population, education, employment and social security.
- Approval from The Norwegian Medicines Agency (Statens legemiddelverk, SLV). E.g. for a clinical drugs trial
- Confidentiality permit (exemption from the duty of confidentiality) from a department or directorate
- Other approval. E.g. from a Data Protection Officer

Processing

Where will the personal data be processed? In the framework of this project, data will be simulated, meaning no personal data will be collected, used, stored or processed.

“Processing” includes any collecting, registering, storing, collating, transferring etc. of data. You must indicate all processing of personal data that will take place in the project.

- Computer belonging to the institution responsible for the project N/A
- Computer owned/operated by the data controller. For example, processing data in a private or communal user area on the institution's server. N/A
- Mobile device belonging to the data controller. Mobile device owned/operated by the data controller. A mobile device can be a laptop, camera, mobile phone etc. N/A
- Physically isolated computer belonging to the data controller. Not connected to other computers or to a network, neither internally nor externally. N/A
- External service or network. Such as providers of cloud storage, online surveys or data storage (such as TSD). Use of an external service or server requires that a data processor agreement is made between the data controller and the external party. N/A
- Private device. Data collection or storage on private devices such as your own computer or mobile phone etc. is not recommended and must be clarified with the institution responsible for the project. N/A

Who will be processing/have access to the collected personal data? N/A

- Project leader N/A
- Student (student project) N/A
- Internal co-workers. Employees of the data controller. N/A
- External co-workers/collaborators inside the EU/EEA. Employees of other institutions that have formalized cooperation with the data controller, or employees of other institutions that are joint data controllers. N/A
- Data processor. An external person or entity that processes personal data on behalf of the data controller, such as an online survey provider, cloud storage provider, translator or transcriber. There must be a data processor agreement or other legal agreement between the data controller and the external party. N/A

- Others with access to the personal data. N/A

Which others will have access to the collected personal data? N/A

Will the collected personal data be made available to a third party or international organisation outside the EEA? This includes when personal data are sent to and stored in a country outside the EEA, or when persons outside this area are given access to personal data stored within the EEA. This means that you cannot use a service provider or outsourced supplier outside the EEA, unless there is a valid basis for the transfer of personal data. Yes No N/A

Give the name of the institution/organisation N/A Give the country of the institution/organisation N/A On what basis will the collected personal data be transferred? N/A

Personal data can be transferred on the basis of an adequate level of protection (art. 45) or on the basis of appropriate safeguards (art. 46). Personal data can also be transferred on the basis of the exception for special situations, but only if the transfer is not repeated, concerns only a limited number of data subjects, is necessary for the purposes of compelling legitimate interests pursued by the data controller (which are not overridden by the interests or rights and freedoms of the data subject), and if the data controller has assessed all the circumstances surrounding the data transfer and has provided suitable safeguards with regard to the protection of personal data (art. 49).

Information Security

No personal data will be used in the project. Therefore, identification and/or security issues are irrelevant for this project.

Will directly identifiable personal data be stored separately from the rest of the collected data (in a scrambling key)?

It is common practice to remove directly identifiable data (name, national ID number, contact details etc.) from the collected data and give each data subject a code/number. A scrambling key is the file/list of names and codes that makes it possible to directly identify data subjects in the collected data. It should be stored separately from the rest of the collected data. In practice, this means that the scrambling key cannot be stored in the same network as the rest of the data, unless the scrambling key is encrypted. Yes No N/A

Explain why directly identifiable personal data will be stored together with the rest of the collected data. N/A

For reasons of information security we recommend the use of a scrambling key in most projects, especially in projects where special categories of personal data (previously "sensitive" personal data) or personal data relating to criminal convictions and offences will be processed.

Which technical and practical measures will be used to secure the personal data?

- Personal data will be anonymized as soon as no longer needed. N/A

  Anonymization involves processing the data in such a way that no individual persons can be identified in the data that you're left with, i.e. the data can no longer be linked to individual persons in any way.

  Anonymization usually involves: *deleting directly identifiable personal data (including scrambling key/list of names) *deleting or rewriting indirectly identifiable personal data (e.g. deleting or categorizing variables such as age, place of residence, school etc.) *deleting or editing audio recordings, photographs and video recordings.

- Personal data will be transferred in encrypted form. N/A

  Encryption is a mathematical method for ensuring confidentiality in that information cannot be read by unauthorized persons. For example, using an encrypted VPN tunnel or equivalent measure for external login to work-place network.

- Personal data will be stored in encrypted form. N/A

  Encryption is a mathematical method for ensuring confidentiality in that information cannot be read by unauthorized persons. For example, the encryption of a hard drive to ensure the confidentiality of data when the computer is turned off.

- Record of changes. N/A

  Changes in the collected data are recorded/documented with the time of the change and information about the person who made that change.

- Multi-factor authentication. N/A

  A method of access control where a user is granted access after presenting two or more separate pieces of evidence to prove their identity (e.g. password + code sent by text message)

- Restricted access. N/A

  Blocking or restricting access to the collected data for unauthorized persons

- Access log. N/A

  An access log shows who has accessed the collected data and when

- Other security measures. N/A

  For example, locking away documents, automatic screen lock after a short time for mobile devices, partitioning of hard drive, checksum/integrity check etc.

Duration of project

Project period

Will personal data be stored beyond the end of project period? Personal data should not be further processed a way that is inconsistent with the initial purpose(s) for which the data were collected. Anonymous/anonymized data may be stored indefinitely, so long as nothing else has been agreed to by the data subjects.

- No, all collected data will be deleted
- No, the collected data will be stored in anonymous form. Stored in a form where the data can no longer be linked to individual persons in any way
- Yes, collected personal data will be stored until
- Yes, collected personal data will be stored indefinitely.

- Other No personal data will be processed. Questions of storing personal sensitive data are irrelevant for this project.

For what purpose(s) will the collected personal data be stored?

- Research
- Other No personal data will be processed. Questions of storing personal sensitive data are irrelevant for this project.

Where will the collected personal data be stored?

- At the institution responsible for the project (data controller)
- Other No personal data will be processed. Questions of storing personal sensitive data are irrelevant for this project.

Additional information

Will the data subjects be identifiable (directly or indirectly) in the thesis/publications for the project? If personal data are to be published, there should be a scientific purpose for this. Data is usually published in anonymous form. Yes No N/A

Explain why N/A

Additional information N/A

Here you can provide information that may have significance for our assessment of the project, including more detailed information about points covered in the form and information that is not covered by points in the form.

Other attachments N/A e.g. interview guide, questionnaire, information letter and consent form etc.

Appendix II

Syntax

As this experiment involved testing and comparison of two different model selection approaches in a variety of different simulated settings, each individual setting corresponds to an individual syntax script. In total, 84 scripts were run in this experiment, and all 84 scripts are available in the attached folder, titled Supplementary Scripts. Scripts are titled with the following structure: [Item Count][Approach][Sample Size][DIF Concentration][Magnitude Setting]. As the experiment included 84 sample scripts, documentation for two sample scripts is provided here; one sample script explains the traditional procedure and the other sample script explains the stepwise procedure. This documentation can be applied to all remaining scripts, depending on whether the traditional or stepwise procedure was run in that script.

The first sample script for which documentation is provided is titled "6 TRAD n500 DIF A," where the traditional procedure was run on the 6-item instrument with 17% DIF, and sample size was 500. The item expressing DIF was Item No. 1 and DIF magnitude setting was setting A, where the reference group discrimination parameter was 0.5 higher than the focal group discrimination parameter.

The second sample script for which documentation is provided is titled, "6 STEP n500 DIF A," where the stepwise procedure was run on the 6-item instrument with 17% DIF, and sample size was 500. The item expressing DIF was item No. 1 and DIF magnitude setting was setting A, where the reference group discrimination parameter was 0.5 higher than the focal group discrimination parameter.

```
#load required packages
install.packages("mirt")
install.packages("data.table")
library(mirt)
library(data.table)
setwd()
getwd()

```

```
# Simulation 2:   6 TRAD n500 DIF A

#set.seed for data-generating function
set.seed(1)
#Item(s) to express DIF, if any. In this case, item expressing DIF is Item No. 1.
ChosenItem <- sample(1:6, 1)
print(ChosenItem) #Item No. 1
#Reference Group Alpha Parameters
refalphas <- runif(6, 0.8, 2.5)
print(refalphas)
#Reference Group Delta Parameters
refdeltas <- runif(6, -2, 2)
print(refdeltas)
#2PL data generating function.
irtresponse2pl <- function(alpha, delta,theta){
  J <- length(alpha)
  N <- length(theta)
  res <- matrix(0, nrow = N, ncol = J)
  for(j in 1:J){
    res[,j]<- runif(N) < (exp(alpha[j] * (theta - delta[j])))/ (1 + exp(alpha[j] * (theta - delta[j])))
  }
  return(res)
}
#define objects for which different results will be stored, once the procedure begins
BICmatrix <- matrix(NA, nrow = 100, ncol = 64)
result.matrix.all <- matrix()
result.DT.all <- data.table()
result.DT.long <- data.table()
#Final results for all 100 replications comprising traditional
```

```
procedures stored in result.DT.Final
result.DT.final <- data.table()
#All combinations of relaxed and constrained items possible for a 6-item instrument.
#0s represent relaxed items
#Non-zero values represent constrained items
Combos=expand.grid(c(0,1),c(0,2),c(0,3),c(0,4),c(0,5),c(0,6))
colnames(Combos) <- c("Item 1", "Item 2", "Item 3", "Item 4","Item 5","Item 6")

#Simulation Begins
for(z in 1:100){
  v1 <- seq(201,300,1) #internal seeds for this simulation
  set.seed(v1[z])
  refthetas <- rnorm(250,0,1) #reference group theta parameter
  focthetas <-rnorm(250,-1,1.2) #focal group theta parameter
  round.alpha <- c(0.5,0,0,0,0,0) #the discrimination parameter for item expressing DIF, if any, is adjusted accordingly
  round.delta <- c(0,0,0,0,0,0) # the difficulty parameter for item expressing DIF, if any, if adjusted accordingly
  focalphas <- refalphas - round.alpha #focal group discrimination parameter
  focdeltas <- refdeltas + round.delta #focal group difficulty parameter
  refdata <- irtresponse2pl(refalphas, refdeltas, refthetas) #reference group's response data is simulated
  colnames(refdata) <- c("Item 1", "Item 2", "Item 3", "Item 4","Item 5","Item 6")
  focdata <- irtresponse2pl(focalphas, focdeltas, focthetas) #focal group's response data is simulated
  colnames(focdata) <- c("Item 1", "Item 2", "Item 3", "Item 4","Item 5","Item 6")
```

```r
final.data <- rbind(refdata,focdata)
#the traditional BIC procedure begins. All combinations of invariance from Combos are assessed, and BIC is assigned to each model:
list1 = list()
for(i in 2:64){ #the first row of the Combos object is not assessed, as this model is fully relaxed and not identifiable
  list1[[i]] = Combos[i,]
}
for(i in 2:64){
  vect_it = unlist(list1[[i]])
  mybase <- try(multipleGroup(final.data, model ='F1 = 1-6', group = c(rep('G1REF', 250), rep('G2FOC', 250)), itemtype=rep('2PL', 6), invariance=c("free_means", "free_var", colnames(final.data)[vect_it]), SE=TRUE, technical=list(NCYCLES=2000)))
  if(class(mybase) == "try-error"){
    BICmatrix[z, i] <- 1
    next
  }
  if(mybase@OptimInfo$converged){
    if(mybase@OptimInfo$converged && mybase@OptimInfo$secondordertest){
      BICmatrix[z, i] <- extract.mirt(mybase, "BIC")}  #store the BIC for the model in question, unless non-convergence occured in which case BIC will be "0" or "1"
  } else{
    BICmatrix[z, i] <- 0
  }
}
#transposing occurs to stack BICs and models accordingly:
result.matrix.all <- cbind(Combos, t(BICmatrix))
result.DT.all <- as.data.table(result.matrix.all)
result.DT.long <- melt(result.DT.all, measure.vars=7:ncol(result.DT.
all), variable.name="trial.z", value.name="BIC.z")
  result.DT.final <- result.DT.long[, .SD[which.min(BIC.z)], by=trial.z] #for each replication, the model with the lowest BIC is known as the "selected model"
}
save.image(file='6TRAD.n500.DIF.A') #upd
#Print BICs corresponding to the 100 selected models for 100 replications (z), for display purposes
table(result.DT.final$BIC.z)
#Manaully tag the true model for this simulation
target <- result.DT.final[3,2:7]
#Which rows have a BIC of zero? These replications (z) must be excluded  from the corresponding stepwise simulation
TradZeroes <- result.DT.final[BIC.z==0.000,]
#Removing non-converging cases from the results table
TradNoNoise <- result.DT.final[BIC.z != 0.000]
#success table reports how many times the procedure in this simulation selected the true model ("CORRECT") versus incorrect model ("NO")
for(d in 1:nrow(TradNoNoise)){
  success <- TradNoNoise[d, Accuracy:= ifelse(isTRUE(all.equal(TradNoNoise[d,2:7], target)),"CORRECT","NO")]
}
table(success$Accuracy) #final report of the number of times the procedure selected the true model accurately (counts of "CORRECT")
######################################################################

#load required packages
install.packages("mirt")
install.packages("data.table")
library(mirt)
library(data.table)

```

```
# Simulation 8:   6 STEP n500 DIF A

#set seed for data-generating function
set.seed(1)
#Item(s) to express DIF, if any. In this case, item expressing DIF is Item No. 1.
ChosenItem <- sample(1:6, 1)
print(ChosenItem) #Item no. 1
#Reference Group Alpha Parameters
refalphas <- runif(6, 0.8, 2.5)
#Reference Group Delta Parameters
refdeltas <- runif(6, -2, 2)
#All combinations of relaxed and constrained items possible for a 6-item instrument.
#0s represent relaxed items
#Non-zero values represent constrained items
Combos=expand.grid(c(0,1),c(0,2),c(0,3),c(0,4),c(0,5),c(0,6))
colnames(Combos) <- c("Item 1", "Item 2", "Item 3", "Item 4","Item 5","Item 6")
# 2PL data generating function
irtresponse2pl <- function(alpha, delta,theta){
  J <- length(alpha)
  N <- length(theta)
  res <- matrix(0, nrow = N, ncol = J)
  for(j in 1:J){
    res[,j]<- runif(N) < (exp(alpha[j] * (theta - delta[j])))/ (1 + exp(alpha[j] * (theta - delta[j])))
  }
  return(res)
}
# z represents the replication number. The stepwise procedure always begins with z=1
```

```
z <- 1
# b parameter relates to the number of items that are relaxed for a given replication. The number of items relaxed for a given replication
# is b or b-1. Thus, in the first replication, the procedure will begin with models where one item, or zero items are relaxed. If b is increased to
# two, then the procedure is assessing models in which 2 items, or 1 item, is relaxed, so on and so forth.
b <- 1
# naming objects where data will be stored
result.matrix <- matrix(NA, nrow = 100, ncol = 7, byrow = TRUE)
colnames(result.matrix) <- c("Item 1", "Item 2", "Item 3", "Item 4","Item 5","Item 6", "BIC")
RLXD.L <-list(mode="vector")
constrained <- list(mode="matrix")
t.constrained <- list(mode="matrix")
trial <- list()
#simulation begins with z=1 and upon selecting a model in the 100th replication (z=100), the function will cease
while (z <= 100) {
  v1 <- seq(201,300,1)  #internal seeds for this simulation
  set.seed(v1[z])
  refthetas <- rnorm(250,0,1) #reference group theta parameter
  focthetas <- rnorm(250,-1,1.2) #focal group theta parameter
  round.alpha <- c(0.5,0,0,0,0,0) #the discrimination parameter for item expressing DIF, if any, is adjusted accordingly
  round.delta <- c(0,0,0,0,0,0) #the difficulty parameter for item expressing DIF, if any, is adjusted accordingly
  focalphas <- refalphas - round.alpha #focal group discrimination parameter
  focdeltas <- refdeltas + round.delta #focal group difficulty
```

```
parameter
refdata <- irtresponse2pl(refalphas, refdeltas, refthetas) #reference group's response data is simulated
colnames(refdata) <- c("Item 1", "Item 2", "Item 3", "Item 4","Item 5","Item 6")
focdata <- irtresponse2pl(focalphas, focdeltas, focthetas) #focal group's response data is simulated
colnames(focdata) <- c("Item 1", "Item 2", "Item 3", "Item 4","Item 5","Item 6")
final.data <- rbind(refdata,focdata)
b <- b #new replications (z) will always begin with b=1
print(b)
print("top b") #if b (related to the number of relaxed items) is increased in the simulation, "top b" will reflect this update. If b is not increased, "top b" will continue to show 1.
print(z)
print ("top z") #serves as a crosscheck to see if the simulation is still working on the same replication, or a new replication has begun
trial[[z]] = list(c(z,b)) #stores the number of relaxations (b) that occured in each replication (z) before a model was selected.
# b = 1 for either the first replication (z=1) or when a new replication has begun
# if b = 1, this means the models to be assessed by the stepwise procedure are those with b items relaxed, or less than b items relaxed
# If b=1, the models being assessed in "Base" are those with 1 or 0 items relaxed
#constrained[[b]] refers to the model with the fewest relaxed items (0) in the models in "Base"
if(b == 1){
 BICmatrix <- matrix(NA, nrow = 8-b, ncol = 1, byrow = TRUE)
```

```
  Base <- matrix(NA, nrow = 8-b, ncol = 6)
  Base <- as.matrix(Combos[rowSums(Combos==0) <= b, ])
  constrained[[b]] <- as.matrix(Base[rowSums(Base==0) < b, ])
}
list1 =list()
#the stepwise BIC procedure begins. Only the models in the "Base" matrix are assessed, and a BIC is assigned to each model in "Base":
for(i in 1:nrow(Base)){
  l = Base[i,]
  list1[[i]] <- l
}
for(i in 1:nrow(Base)){ #upd N
  mybase <- try(multipleGroup(final.data, model ='F1 = 1-6', group = c(rep('G1REF', 250), rep('G2FOC', 250)), itemtype=rep('2PL', 6), invariance=c("free_means", "free_var", colnames(final.data)[list1[[i]]]), SE=TRUE, technical=list(NCYCLES=2000)))
  if(class(mybase) == "try-error"){
    BICmatrix[i,] <- 1
    next
  }
  if(mybase@OptimInfo$converged){
    if(mybase@OptimInfo$converged && mybase@OptimInfo$secondordertest){
      BICmatrix[i,] <- extract.mirt(mybase, "BIC")
    }
  }else{
    BICmatrix[i,] <- 0
  }
}
Results.Base <- cbind(Base,BICmatrix[1:(8-b),])  #results from the function are stored here
#transposing to stack BIC alongside models extracted from function
```

```
 above
Results.MIN <- as.matrix(Results.Base[which.min(Results.Base[,7]), ])
t.Results.MIN <- t(Results.MIN)
t.constrained[[b]] <- t(constrained[[b]])
#Identity test: The sequence of relxed items in the most constrained model in "Base" is already known. An identity test is performed
#to see if the model with the lowest BIC extracted from "Base" has the same sequence of relaxed items as the most constrained model.
# if the identity test is true, and the model with the lowest BIC is in fact the most constrained model, a new replication will begin, "top z" will increase by 1, and "top b" will equal 1. The process begins anew.
# if the identntity test is false, and the model with the lowest BIC is NOT the most constrained model, a new replication does not begin.  z will remain z ("top z" = z), but b will increase by 1 ("top b will be b+1).
#In the latter's case, the stepwise procedure will continue to work on the same replication until the model with the lowest BIC is the most constrained model. Only then will a new replication begin.
IDTest <- all.equal(as.vector(t.constrained[[b]]), as.vector(t.Results.MIN[,1:6]),check.attributes=FALSE,use.names=FALSE)
if(isTRUE(IDTest)){
  result.matrix[z,] <- t.Results.MIN
  b <- 1
  z = z+1
}else{
  z <- z
  b = b+1
  #Base <- matrix(NA, nrow = 8-b, ncol = 6)
  BICmatrix <- matrix(NA, nrow = 8-b, ncol = 1, byrow = TRUE)
  RLXD.L[[b]] <- as.vector(apply(as.data.frame(t.Results.MIN), 1, function(x) paste(colnames(as.data.frame(t.Results.MIN))[which(x==0)
])))
  print(RLXD.L[[b]])
  print(Base)
  Base <- as.matrix(Combos[rowSums(Combos==0) >=(b-1), ])
  Base <- as.matrix(Base[rowSums(Base==0) <= b,])
  Base <- Base[which(apply(as.matrix(Base[, RLXD.L[[b]]]), 1, sum) == 0),]
  constrained[[b]] <- as.matrix(Base[rowSums(Base==0) < b, ])
 }
}
save.image(file=’6STEP.n500.DIF.A’) #store the environment for above code in working directory.
result.DT.final <- as.data.table(result.matrix)
table(result.DT.final$BIC) #Print BICs corresponding to the 100 selected models for 100 replications (z), for display purposes
# Successes
target <- result.DT.final[3,1:6] #manually tag the model that is the true model
StepNoNoise <- result.DT.final[-c(10,15,35,38,48,69,73,85),] #remove the simulations that did not converge in either this simulation or the corresponding simulation for the traditional procedure.
# Successes
for(d in 1:nrow(StepNoNoise)){
 success <- StepNoNoise[d, Accuracy:= ifelse(isTRUE(all.equal(StepNoNoise[d,1:6], target)),"CORRECT","NO")]
}
table(success$Accuracy)

#final report of the number of times the procedure selected the true model accurately (counts of "CORRECT")
#Both simulations’ environments can be pulled up through:
load(’6TRAD.n500.DIF.A’)
load(’6STEP.n500.DIF.A’)
```