UMEÅ UNIVERSITY

Extensions and Applications of Item Response Theory

Joakim Wallmark

Department of Statistics
Umeå School of Business, Economics and Statistics
Umeå, 2025

Doctoral Thesis
Department of Statistics
Umeå School of Business, Economics and Statistics
Umeå University
SE-901 87 Umeå

Statistical Studies No. 60
ISBN: 978-91-8070-572-1 (electronic)
ISBN: 978-91-8070-571-4 (print)
ISSN: 1100-8989
Electronic version available at http://umu.diva-portal.org/

Printed by: UmU Print Service, Umeå University
Umeå, Sweden 2025

Contents

List of Papers

The following papers are included in the thesis:

I. Wallmark, J., Josefsson, M., and Wiberg, M. (2023). Efficiency analysis of item response theory kernel equating for mixed-format tests. *Applied Psychological Measurement*. https://doi.org/10.1177/01466216231209757

II. Wallmark, J., Ramsay, J. O., Li, J., and Wiberg, M. (2023). Analyzing polytomous test data: A comparison between an information-based IRT model and the generalized partial credit model. *Journal of Educational and Behavioral Statistics*. https://doi.org/10.3102/10769986231207879

III. Wallmark, J., Josefsson, M., and Wiberg, M. (2024). Introducing flexible monotone multiple choice item response theory models and bit scales. *Submitted Manuscript*. Preprint available at https://arxiv.org/abs/2410.01480

IV. Wallmark, J. (2024). IRTorch: An item response theory package for Python. *Submitted Manuscript*. Package available at https://github.com/joakimwallmark/irtorch

Abstract

This doctoral thesis focuses on Item Response Theory (IRT), a statistical method widely used in fields such as education and psychology to analyze response patterns on tests and surveys. In practice, IRT models are estimated using collected test data, which allows researchers to assess both how effectively each item measures the underlying trait—such as subject knowledge or personality characteristics—that the test aims to evaluate, and to estimate each individual's level of that trait. Unlike traditional methods that simply sum predetermined item scores, IRT accounts for the difficulty of each item and its ability to measure the intended trait.

The thesis consists of four research articles, each addressing different aspects of IRT and its applications. The first article focuses on test equating, ensuring that scores from different versions of a test are comparable. Equating methods with and without IRT are compared using simulations to explore the advantages and disadvantages of incorporating IRT into the kernel equating framework. The second and third articles introduce and compare different types of IRT models. Through simulations and real test data examples, these studies demonstrate that more flexible models can better capture the true relationships between test responses and the underlying traits being measured.

Finally, the **IRTorch** Python package is presented in the fourth study. **IRTorch** supports various IRT models and estimation methods and can be used to analyze data from different types of tests and surveys. In summary, the thesis demonstrates how IRT-based equating methods can serve as an alternative to traditional equating methods, how more flexible IRT models can improve the precision of test results, and how user-friendly software can make advanced statistical models accessible to a wider audience.

KEYWORDS: Item response theory, Test equating, Statistical software, Educational assessment, Latent variable modelling

Populärvetenskaplig sammanfattning

Avhandlingen fokuserar på Item Response Theory (IRT), en statistisk metod som främst används inom utbildning och psykologi för att analysera data från prov och enkäter. I praktiken estimeras IRT-modeller med hjälp av insamlade data och används sedan för att bedöma hur bra varje uppgift mäter den egenskap som provet syftar till att mäta (t.ex. ämneskunskap eller personlighetsdrag). Modellerna kan även användas för att uppskatta varje individs nivå av den egenskapen. Istället för att bara summera förbestämda uppgiftspoäng för varje uppgift tar IRT hänsyn till hur svår varje uppgift är, och hur bra den mäter det som provet är avsett att mäta.

Avhandlingen består av fyra forskningsartiklar som var och en belyser olika aspekter av IRT och dess tillämpningar. I den första behandlas provekvivalering, som säkerställer att poäng från olika versioner av ett prov är jämförbara. Genom ekvivalering kan skillnader i svårighetsgrad mellan olika provversioner justeras. Ekvivaleringsmetoder med och utan IRT jämförs med hjälp av datorsimuleringar för att undersöka för- och nackdelar med att använda IRT som en del av ekvivaleringsprocessen.

Den andra och tredje artikeln introducerar och jämför olika typer av IRT-modeller. Med hjälp av simuleringar och exempel med verkliga provdata illustreras att mer flexibla modeller bättre kan fånga de verkliga sambanden mellan provsvar och de underliggande egenskaperna som mäts.

Slutligen presenteras ett Python-paket, **IRTorch**, som utvecklats för att underlätta användningen av IRT i praktiska tillämpningar. **IRTorch** stödjer flera olika IRT-modeller och skattningsmetoder, och kan användas för att analysera data från olika typer av prov.

Sammanfattningsvis visar avhandlingen hur IRT-baserade ekvivaleringsmetoder kan användas som alternativ till traditionella ekvivaleringsmetoder, hur mer flexibla IRT-modeller kan användas för att förbättra precisionen i provresultat, och hur användarvänlig programvara kan göra avancerade statistiska modeller tillgängliga för fler användare.

Acknowledgements

I would like to thank my supervisors, Marie and Maria, for their invaluable guidance, support, and feedback. You have both provided tremendous help, often going above and beyond what was expected. I would also like to acknowledge all my co-authors and colleagues for sharing this academic journey. A special thanks goes to Filip, whose engaging discussions on research and statistical methodologies enriched my experience and understanding. Lastly, I am grateful to my girlfriend, Emilija, for her steady support and patience during this process.

Umeå, December 2025
Joakim Wallmark

1 Introduction

Psychometrics involves assessing latent traits that cannot be directly observed (e.g., knowledge, intelligence, personality traits) through observable indicators like test or questionnaire items (e.g., Sijtsma and van der Ark, 2020, pp. 5-6). A key analytical framework in this field is Item Response Theory (IRT, Birnbaum, 1968), which has become a cornerstone of modern educational and psychological assessment. IRT models aim to capture the relationships between the latent trait and the possible responses of each test item. Unlike earlier methods, where a test score is simply calculated as the sum or average of the item scores, IRT provides a more precise and flexible way of understanding and analyzing data from an assessment.

Although IRT has been widely adopted, several challenges and opportunities for further development remain. This thesis focuses on addressing some of these gaps through four interconnected studies that explore both theoretical and practical aspects of IRT. These studies tackle methodological issues and introduce new tools, along with supporting software for IRT analysis. The goal of this work is to enhance IRT's utility, contributing to more accurate and fair assessments in educational and psychological measurement.

The thesis is structured as follows: Section 2 provides a brief overview of different test formats. Section 3 provides an overview of IRT, its foundational principles, and key assumptions. It also introduces many of the IRT models relevant to the studies in the thesis. Section 4 discusses the estimation algorithms used in the studies, including marginal maximum likelihood, joint maximum likelihood, and autoencoder neural networks. Section 5 outlines methods for latent variable estimation using a fitted IRT model. Section 6 discusses the arbitrary nature of the latent variable scale in IRT models and the implications of this for model interpretation. Section 7 presents the real datasets used throughout the studies in the thesis: the Swedish Scholastic Aptitude Test (SAT) and the national test in mathematics for Swedish high school students. Section 8 summarizes the four papers included in the thesis, highlighting their key findings and contributions. Finally, Section 9

offers concluding remarks and suggests avenues for future research.

2 Items and test forms

Various types of test forms are used in educational and psychological assessments, each with its own unique characteristics and challenges. Some of the more common types include multiple-choice tests, constructed-response tests, and mixed-format tests. All of these can be analyzed using IRT, but different models may be more or less suitable depending on the test format.

Multiple-choice tests consist of questions with a set of predefined answers, and test-takers must choose the correct one. These tests are easy to score because answers are either right or wrong, making them ideal for large-scale assessments. One limitation is that they might encourage guessing, and they do not typically assess deeper understanding or critical thinking (Scouller, 1998).

Constructed-response tests require test-takers to generate their own answer to each item. Items are typically scored on an ordered scale, where the score is determined by the quality of the response. Examples include essay questions or math problems where the solution must be explained. These tests are better suited for evaluating higher-order thinking skills, such as analysis and problem-solving, because they allow students to show their reasoning and creativity (Birenbaum and Feldman, 1998). However, scoring is more complex and time-consuming, often involving human graders or detailed rubrics, which can introduce subjectivity.

By including items with different scoring formats in a single test, one can mitigate the limitations of each format while leveraging their strengths. Such tests are typically referred to as mixed-format tests (Kim et al., 2010; Kolen and Lee, 2018). Both multiple-choice and constructed-response items may be included in the same mixed-format test.

Often, multiple versions of a test are administered at different times. Examples of such tests include the Swedish SAT and the national test in mathematics for Swedish high school students, both described in more detail in Sections 7.1-7.2. To ensure that scores

from different versions are comparable, test equating is performed (Wiberg et al., 2025). This process adjusts for differences in difficulty between test versions, allowing for fairer comparisons of test-taker performance. The first study in this thesis focuses on test equating, comparing equating methods with and without IRT.

3 Item response theory

IRT is based on the idea that the probability for a test taker to respond with a particular response to an item can be explained by the latent trait measured by the test or questionnaire (Birnbaum, 1968; Lord and Novick, 1968). IRT models represent these relationships for all items through item response functions (IRFs). In this section, some of the most common modelling assumptions in IRT are presented together with various IRT models.

Consider a test with items indexed by $j = 1, 2, \ldots, J$ and test takers indexed by $i = 1, 2, \ldots, N$. The response pattern for test taker i is represented by the random vector $\mathbf{X}_i = (X_{i1}, X_{i2}, \ldots, X_{iJ})$, where each element X_{ij} indicates the response by test taker i on item j. The observed responses are denoted by $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iJ})$. When the identity of the test taker is redundant, the subscript i is omitted, and $\mathbf{X} = (X_1, X_2, \ldots, X_J)$ refers to the response vector for a generic test taker. Furthermore, let $m = 0, 1, ..., M_j$ denote the possible responses to item j. From here, the IRF for response m on item j is $P(X_j = m \mid \theta)$, where θ is the latent variable measured by the item.

3.1 Model assumptions

The main assumptions of the IRT models considered in this thesis are unidimensionality, local independence, monotonicity, and that the model can closely approximate the true data generating process (Hambleton and Swaminathan, 1985, Chapter 2).

Unidimensionality: The latent variable is assumed to be unidimensional, meaning that it can be represented by a single continuous variable. In practice, the unidimensionality assumption

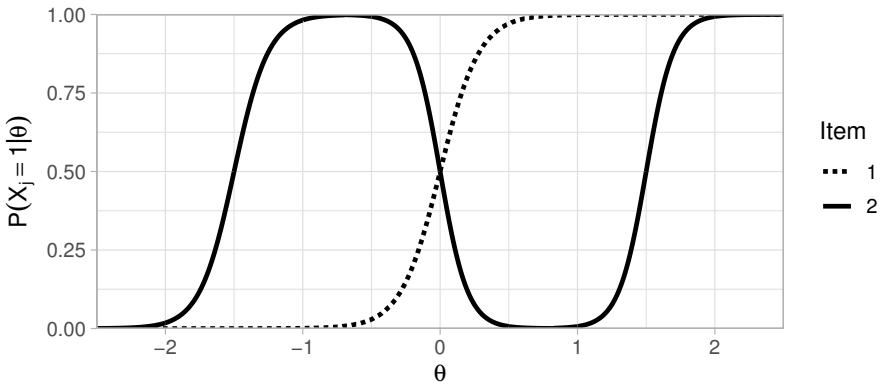

Figure 1: Example IRFs without assuming monotonicity.

is often tested using factor analysis or other statistical methods. See Chapter 3.4 in Martinková and Hladká (2023) for some practical examples. One should note that while many models have been extended to accommodate multidimensional latent variables, the focus of the thesis is primarily on unidimensional models.

Local independence: The local independence assumption states that the item responses are conditionally independent given the latent variable θ. This means that the response probabilities of one item do not depend on the responses to other items after accounting for the latent variable. Violations of local independence can lead to biased parameter estimates and inaccurate inferences.

Monotonicity: The IRF is often assumed to be monotonically increasing in θ. This means that as the latent variable increases, the probability of a correct response/higher score on an item also increases. Violations of monotonicity can lead to difficulties in interpreting the model. To illustrate, consider a test with only two dichotomously scored items that measure something vastly different. Maybe item one contains a question about the history of the Roman Empire, while item two contains a question about linear algebra. If we allow for the IRFs to freely oscillate up and down, we could end up with something like the plot in Figure 1. The curves can capture any response pattern perfectly, but the interpretation of

the latent variable is difficult. Does θ represent math or history? If it is math, then why does the probability of a correct response to the history item change as θ changes? We essentially incorporate multiple latent variables into one variable. This example could be easily extended to tests or questionnaires with more items. The IRFs would simply have more oscillations, and still capture every possible response pattern perfectly. Among the models considered in this thesis, the OS model, introduced in Section 3.4, is the only one that does not assume monotonicity.

The model closely approximates the true data-generating process: This is a general assumption that is pretty much always made when fitting a statistical model. It is assumed that the model is a good approximation of the true data generating process. In other words, the IRFs must be sufficiently flexible without overfitting the training data.

3.2 Traditional parametric models

The most common IRT models for dichotomously scored items (items with only two possible responses) are the Rasch model (Rasch, 1960), the two-parameter logistic (2PL) model (Birnbaum, 1968), and the three-parameter logistic (3PL) model (Birnbaum, 1968). The IRF for the 3PL model is defined as

$$P(X_j = 1 \mid \theta) = c_j + \frac{1 - c_j}{1 + \exp(-a_j(\theta - b_j))}, \tag{1}$$

where a_j, b_j, and c_j are item parameters associated with item j. The parameter a_j is the discrimination parameter, indicating the item's ability to differentiate between individuals with different levels of the latent variable. The parameter b_j is the difficulty parameter, where a larger b_j indicates a more difficult item. The parameter c_j is the guessing parameter, indicating the probability of a correct response when the latent variable is at its lowest level. The 2PL model is a special case of the 3PL model where the guessing parameter is set to zero, $c_j = 0$. The Rasch model is a further simplification of the 2PL model where the discrimination parameter is set to one across all items, $a_j = 1$.

For polytomously scored items (items with more than two possible responses), the Generalized Partial Credit (GPC) model (Muraki, 1992) and the Nominal Response (NR) model (Bock, 1972) are commonly used. The GPC model is designed for items with ordered responses, such as constructed-response test items, and its IRF is defined as

$$P(X_j = m \mid \theta) = \begin{cases} \dfrac{1}{1 + \sum_{g=1}^{M_j} \exp\left(g a_j \theta - \sum_{t=1}^{g} b_{jt}\right)}, & \text{if } m = 0 \\ \dfrac{\exp\left(m a_j \theta - \sum_{t=1}^{m} b_{jt}\right)}{1 + \sum_{g=1}^{M_j} \exp\left(g a_j \theta - \sum_{t=1}^{g} b_{jt}\right)}, & \text{otherwise.} \end{cases} \tag{2}$$

Like the 2PL and 3PL models, the GPC model has a single discrimination parameter per item a_j. The difficulty parameters b_{jt} define the points on the latent continuum at which the probability of responding with a score of t surpasses that of responding with $t - 1$. Note that when $M_j = 1$, the GPC model reduces to the 2PL model.

The NR model allows for items with unordered response categories. It defines the IRF for responding with m on item j as

$$P(X_j = m \mid \theta) = \frac{\exp\left(a_{jm}\theta + b_{jm}\right)}{\sum_{t=0}^{M_j} \exp\left(a_{jt}\theta + b_{jt}\right)}, \tag{3}$$

where a_{jm} and b_{jm} are item parameters associated with response m. Here, a_{jm} indicates the discrimination power of item response m between individuals with different levels of the latent variable, and the b_{jm} are location parameters, one for each item response. For a more comprehensive overview of traditional IRT models, see van der Linden (2016).

3.3 Monotone multiple choice model

The MMC model is specifically tailored for multiple choice items. It is one of the main contributions of the third study in this thesis and allows for flexible, non-linear relationships between the latent variable and the item responses. The MMC model can be seen as

an extension of the NR model, where incorrect responses are taken into consideration for test taker scoring. In the MMC model, the linear predictors in the exponents of Equation 3 are replaced with non-linear monotone functions of the latent variable, thus allowing for more flexible curves

$$P(X_j = m \mid \theta) = \frac{\exp\left(z_{jm}(\theta)\right)}{\sum_{t=1}^{M_j} \exp\left(z_{jt}(\theta)\right)}. \tag{4}$$

For each incorrect response option, $z_{jm}(\theta) = \tau_j \delta_{jm}(\theta) + b_{jm}$, where $\delta_{jm}(\theta)$ is any monotone function of θ. To maintain the IRT assumption of monotonicity, it is required that $z_{jm}(\theta)$ for the correct response always increases faster than those for the incorrect options. The parameters τ_j allow some items to be negatively correlated with the latent variable. At the time of writing, the model is only available in the **IRTorch** package presented in the fourth study.

3.4 Optimal scoring model

The Optimal Scoring (OS, Ramsay et al., 2020) IRF is defined as

$$P(X_j = m \mid \theta) = \frac{1}{M_j^{S_{jm}(\theta)}},$$

where $S_{jm}(\theta)$ is a continuous function of arbitrary flexibility associated with response m on item j. In the **TestGardener** R package (Ramsay and Li, 2024), the OS model is fitted using a version of joint maximum likelihood (JML, Birnbaum, 1968; Baker and Kim, 2004, Chapter 4), described in Section 4.2. In the package, B-spline functions (Ramsay and Silverman, 2005) are used for $S_{jm}(\theta)$. B-splines are piecewise polynomial functions that can approximate any continuous function. More details on the OS model can be found in the second study.

4 Model estimation algorithms

Brief overviews of the model estimation algorithms considered in the four studies are provided in this section. The likelihood function

is crucial for all of them. Assuming a set of item parameters $\boldsymbol{\xi} = \{\xi_1, \xi_2, \ldots, \xi_J\}$ and latent variables $\boldsymbol{\theta} = \{\theta_1, \theta_2, \ldots, \theta_N\}$, the likelihood function is given by

$$\mathcal{L}(\boldsymbol{\xi}, \boldsymbol{\theta}) = \prod_{i=1}^{N} \prod_{j=1}^{J} P(\mathbf{X}_{ij} = \mathbf{x}_{ij} \mid \xi_j, \theta_i). \tag{5}$$

Note that defining the likelihood function in this way implicitly assumes local independence, as stated in Section 3.1. By taking the logarithm of Equation 5, we obtain the log-likelihood function, which is mathematically more tractable and often used in practice for model estimation

$$\ell(\boldsymbol{\xi}, \boldsymbol{\theta}) = \sum_{i=1}^{N} \sum_{j=1}^{J} \log P(\mathbf{X}_{ij} = \mathbf{x}_{ij} \mid \xi_j, \theta_i). \tag{6}$$

4.1 Marginal maximum likelihood

With marginal maximum likelihood (MML, Bock and Aitkin, 1981), a distribution $g(\theta)$ of the latent variable is assumed. The likelihood function in Equation 5 is then integrated over θ to obtain the marginal likelihood function

$$\mathcal{L}(\boldsymbol{\xi}) = \int \mathcal{L}(\boldsymbol{\xi}, \boldsymbol{\theta}) g(\theta) d\theta. \tag{7}$$

In practice, the integral is often approximated using the EM-algorithm (Bock and Aitkin, 1981), after which the item parameters $\boldsymbol{\xi}$ can be estimated by maximizing the marginal likelihood function. $\boldsymbol{\theta}$ can then be estimated in a second step, as described in Section 5.

4.2 Joint maximum likelihood

The idea behind the JML method is to concurrently estimate both latent variables $\boldsymbol{\theta}$ and item parameters $\boldsymbol{\xi}$ by maximizing the log-likelihood in Equation 6 iteratively (Baker and Kim, 2004, Chapter 4). This process usually begins by initially estimating the item parameters through a heuristic approach. The likelihood is then

maximized with respect to the latent variables $\boldsymbol{\theta}$ while treating the item parameters $\boldsymbol{\xi}$ as constants. In the next step, the roles are reversed: $\boldsymbol{\theta}$ are fixed, and $\boldsymbol{\xi}$ are adjusted to maximize the likelihood.

The algorithm for fitting OS models implemented in the **TestGardener** R package (Ramsay and Li, 2024) is a special case of JML, where the latent variables are constrained within a prespecified range, and the IRFs are estimated using splines (Ramsay and Silverman, 2005).

4.3 Autoencoder neural networks

An autoencoder (AE, Kramer, 1991) is a form of artificial neural network designed to learn efficient representations of input data, mapping it to a lower-dimensional latent space. It can be seen as a non-linear generalization of principal component analysis. An AE comprises two key components: the encoder and the decoder, as illustrated in the top plot of Figure 2. The encoder's role is to compress the input data into a more compact form, commonly referred to as the latent space or bottleneck. The decoder then takes this condensed representation and reconstructs the original input data.

In its basic form, the entire network is a sequence of fully connected hidden layers. Each node in these layers (the gray circles in Figure 2), referred to as neurons, applies a non-linear transformation to a linear combination of the input it receives and passes the result to the next layer. Stochastic gradient descent methods with different optimizers, such as AMSGrad (Reddi et al., 2019), are typically used to find the parameters of the AE by minimizing a loss function based on the difference between the input and the output. The flexibility of the AE can be increased by adding more hidden layers or by increasing the number of neurons within these layers. To fit an IRT model, the entire decoder network is replaced with the IRFs (Curi et al., 2019), as seen in the bottom plot of Figure 2. The negative log-likelihood function in Equation 6 can then be used as the loss function.

In a sense, latent variables $\boldsymbol{\theta}$ and item parameters $\boldsymbol{\xi}$ are estimated

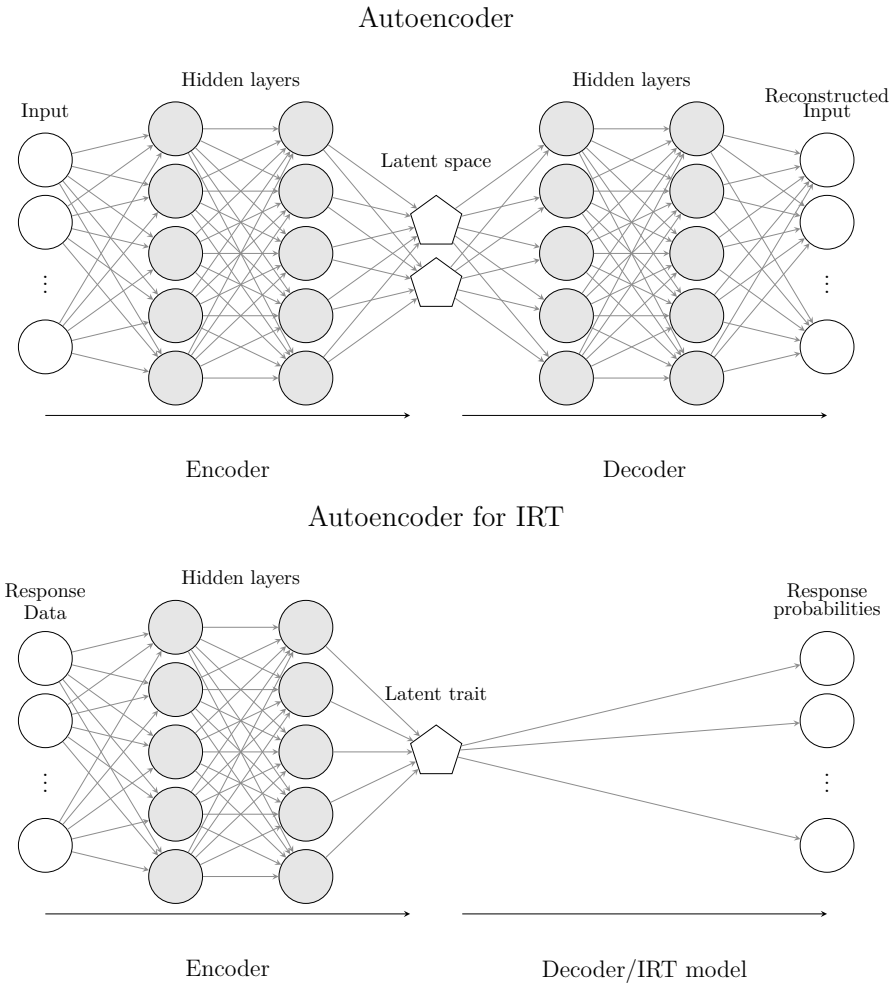

Figure 2: Example of a traditional autoencoder neural network with a two-dimensional latent space, as well as an autoencoder designed for IRT.

simultaneously when an AE is used for model estimation, similar to JML. The AE method also has an inherent advantage in that θ scores of each test taker do not have to be estimated over and over during the fitting process, but instead it learns a function (the encoder) that directly maps a set of observed item responses to the θ scale. This makes it more computationally efficient for large samples or for models with more than one latent variable. Autoencoder models are explained in more detail in the third and fourth study of this thesis. The **IRTorch** package presented in the fourth study also supports AE extensions which allow for Bayesian latent variable inference.

5 Latent variable estimation methods

For a fitted IRT model, there are various ways to estimate the latent variables. Some of the more common methods are maximum likelihood (ML, Baker and Kim, 2004, Chapter 3.2), maximum a-posteriori (MAP, Baker and Kim, 2004, Chapter 7.5.1), and expected a-posteriori (EAP, Baker and Kim, 2004, Chapter 7.5.2). The ML estimate is the value of θ that maximizes the log-likelihood function in Equation 6 given the estimated item parameters.

MAP and EAP are Bayesian methods that incorporate a prior distribution for the latent variables $g(\theta)$ and then utilize the posterior distribution of θ to estimate θ given the observed responses. The EAP estimate is the expected value of the posterior distribution, while the MAP estimate is the mode of the posterior distribution. MAP and EAP estimates reduce the likelihood of extreme θ scores because the prior acts as a regularizer, pulling estimates towards more probable values, reducing uncertainty for smaller samples. However, the choice of prior distribution is crucial, as it can have a significant impact on the estimates. Typically, $g(\theta)$ is a standard normal distribution, but this may not be reasonable for models fitted with no specific assumptions about the latent variable scale, such as with AEs or JML.

For the AE fitted models discussed in Section 4.3, the latent variables can also be estimated by simply passing the test responses

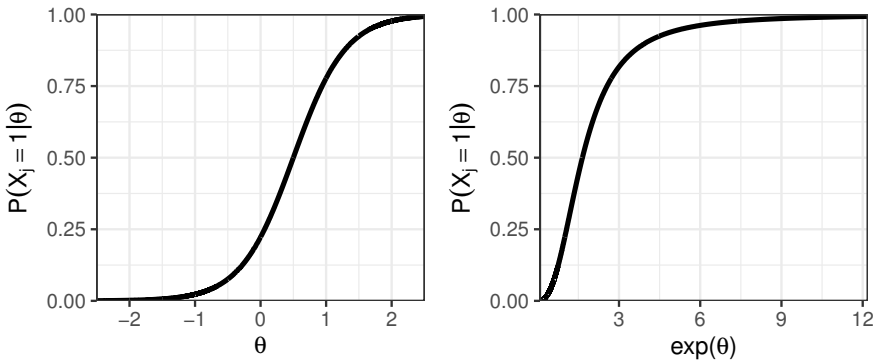

Figure 3: The IRF of a fitted 2PL item. On the left, it is plotted on the original latent variable scale. The same IRF is plotted on the right, but the latent variable scale has been exponentiated.

through the encoder. This is the most computationally efficient approach, as no further optimization is required. If the encoder is well-trained with the log-likelihood as the loss function, then the encoder θ estimates will be close to the ML estimates. The encoder method is discussed and evaluated in the third study of this thesis.

6 Latent variable scales

The latent variable scale of an IRT model is arbitrary and is often assumed to be continuous and normally distributed, both for interpretability and for the estimation algorithm's convenience. Examples of estimation algorithms built around these assumptions include MML (Section 4.1) and Bayesian methods (Levy and Mislevy, 2017). Since the scale is arbitrary, it can be freely transformed as long as the order of the test takers is preserved. An example of this is shown in Figure 3, where the original latent variable scale is exponentiated. The scale has changed, but it is an equally valid model.

Due to this invariance, no scale transformation is inherently superior to another in terms of model fit. However, they may

lead to vastly different interpretations of the latent variable. This is especially true for measures dealing with derivatives/slopes of the latent variable scales, such as discrimination parameters. For example, a_j in models such as the 2PL and 3PL (Equation 1), or measures like Fisher information about θ. For example, it is well established that an ML estimator for θ is asymptotically normal with variance $I(\theta)^{-1}$ (Baker and Kim, 2004; van der Vaart, 2000), where $I(\theta)$ represents the Fisher information

$$I(\theta) = \mathrm{E}_{\mathbf{X}}\left[\left(\frac{\partial \ell(\theta|\mathbf{X})}{\partial \theta}\right)^2\right] = -\mathrm{E}_{\mathbf{X}}\left[\frac{\partial^2 \ell(\theta|\mathbf{X})}{\partial \theta^2}\right].$$

Psychometricians often use such standard errors as measures of precision of the latent variable estimates. While these asymptotic properties can be used to determine, for example, if one test taker likely has a larger true θ than another, the magnitude of the standard error holds little value. To explore this further, we consider a 2PL IRT model fitted using MML to the quantitative section of the Swedish SAT dataset (later introduced in Section 7). The fitting algorithm will pull the θ distribution towards a standard normal, as shown in the upper left plot in Figure 4. However, if we transform the θ scale using for instance, a sigmoid function $1/(1+\exp(-\theta))$, the θ distribution will be more spread out, as seen in the upper right plot in Figure 4. The model is still the same, and no distribution is necessarily better than the other, but the interpretation of their Fisher information curves in the lower plots in Figure 4 (commonly referred to as test information curves in the field of IRT) are vastly different. On the original scale, the information curve suggests that the test measures best for test takers in the middle of the latent variable distribution, while on the transformed scale, the test measures better at the extremes. This is because rates of change on arbitrary scales with no unit of measurement have no meaning, and it has been a long-standing issue in IRT (Lord, 1975).

In the second, third and fourth studies, properties from information theory are used to create scale transformations that have an inherent unit of measurement, reducing the arbitrary nature of the

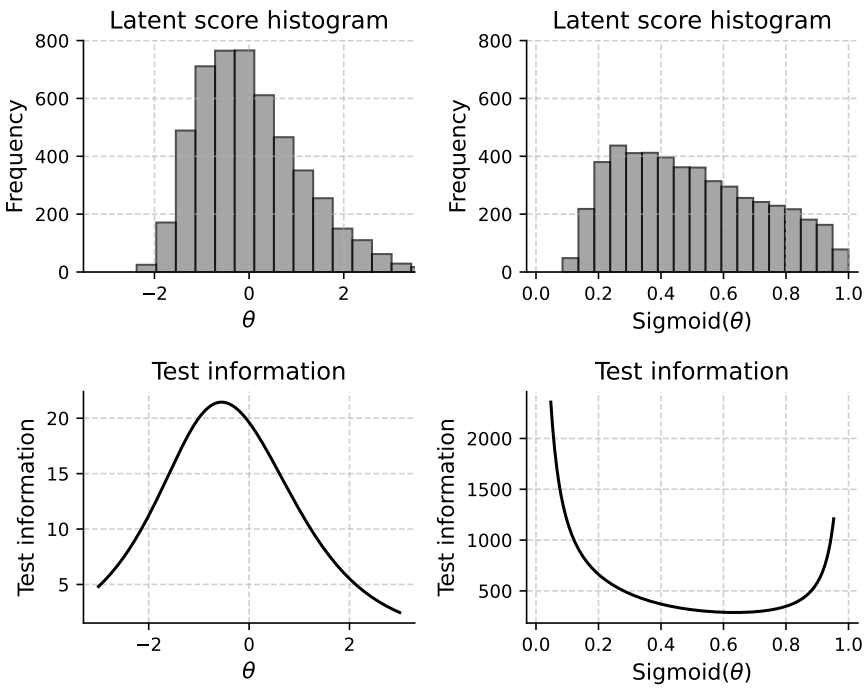

Figure 4: θ distributions and test information curves before and after sigmoid transformation of θ scale.

latent variable scales. These transformations are also presented in the following subsection.

6.1 Information theory for latent variable scales

Information theory is a branch of mathematics that deals with the quantification of information (Shannon, 1948). One of the key concepts in information theory is surprisal, sometimes referred to as self-information, which is the negative logarithm of the probability of a random event. Surprisal measures the information contained within the event and ranges from zero to infinity. It is used in many central statistical concepts, such as the log-odds transform, negative log-likelihood, and KL divergence (Kullback, 1978). Surprisal originates from digital storage of information, where it is used to measure the efficiency of encoding information. Different bases logarithms can be used for computing surprisal, and the resulting unit of measurement depends on the base. When a base-2 logarithm is used, the unit of surprisal is the bit, a binary digit that can only take the values 0 or 1. Throughout this thesis, we use the term Q-bit to refer to the unit of measurement for surprisal computed with a base Q logarithm.

Why these bit units? Let us consider a simple toy example to illustrate this. Suppose that each week, we want to send a summary of our mood each morning over the past week to a friend. We want to do it efficiently using as few bits as possible on average. If we are happy 50% of the time, sad 25% of the time, and angry 25% of the time, we should use fewer bits to encode the happy mood than the sad or angry mood, since we will send it more often. We can evaluate the surprisals for each mood as $-\log_2(0.5) = 1$, $-\log_2(0.25) = 2$, and $-\log_2(0.25) = 2$ bits, respectively. This means the most efficient encoding would be to send only one bit for happy, but two for sad and angry. For example, we could encode happy as '0,' sad as '10,' and angry as '11.'. On the other hand, if we are always happy 100% of the time, we have surprisal $-\log_2(1) = 0$, and thus we don't need to send any bits. Our friend already knows we are happy.

As opposed to probability which has an upper bound of one, surprisal has metric properties and is additive. In the context of IRT, this property can be utilized to take an arbitrary latent variable scale of a fitted model and convert it into a metric scale with the bit unit. This was first done by Ramsay et al. (2020). Since the IRFs are probability curves, they have corresponding surprisal curves $S_Q(X_j = m \mid \theta) = -\log_Q P(X_j = m \mid \theta)$. For a given test, these surprisal curves collectively form a continuous one-dimensional curve in multidimensional space (one dimension per item response category), which is parameterized by θ. The distance along such curve when moving from one θ to another is called the arc length. The arc length from the smallest possible θ, denoted θ_0, to a test taker's estimated θ, denoted $\hat{\theta}$, can be used as a measure of the latent variable, as it is a monotone transformation of the arbitrary θ scale. This IRT surprisal based arc length is computed as

$$A(\hat{\theta}) = \int_{\theta=\theta_0}^{\hat{\theta}} \sqrt{\sum_{j=1}^{J} \sum_{m=0}^{M_i} \left[\frac{dS_{M_j}(X_j = m|\theta)}{d\theta} \right]^2} d\theta. \tag{8}$$

Note that the surprisal base in Equation 8 is the number of possible item responses M_j, as introduced by Ramsay et al. (2020). This means the surprisal curves are measured in M_j-bits, and that $A(\hat{\theta})$ also has the M_j-bit unit. The second study showcases how $A(\hat{\theta})$ can be utilized to bring the latent variable scales of vastly different IRT models onto a common metric scale.

However, $A(\hat{\theta})$ has some limitations. For example, it is not defined when the probability of a correct response is zero, as the surprisal goes to infinity. This is a problem for IRT models that produce very small or even zero probabilities, like most parametric models. Because of this, it makes more sense to use the expected surprisal, typically referred to as *entropy*. For a single item, the entropy curve using base Q surprisal is defined as:

$$
\begin{aligned}
H_{jQ}(\theta) &= \mathbb{E}\left[S_Q(X_j = m \mid \theta)\right] \\
&= \sum_{j=1}^{J} \sum_{m=0}^{M_j} P(X_j = m \mid \theta)\, S_Q(X_j = m \mid \theta).
\end{aligned}
$$

In the third study, an alternative scale transformation to $A(\hat{\theta})$ is introduced, which is based on $H_{jQ}(\theta)$. We refer to the scales resulting from this alternative transformation as *bit scales*, and the scores on the scale are computed as

$$
B(\hat{\theta}) = \sum_{j=1}^{J} \int_{\theta=\theta^{(0)}}^{\hat{\theta}} \left| \frac{dH_{j2}(\theta)}{d\theta} \right| d\theta. \tag{9}
$$

$B(\hat{\theta})$ holds two main advantages over $A(\hat{\theta})$. First, it is defined for all θ values, as the entropy is always finite. Second, calculating a test taker's position on $B(\hat{\theta})$ is more straightforward—it involves simply adding together bit scores from each item. This means that replacing one item with another in the test results in a predictable, constant change in $B(\hat{\theta})$, regardless of the other items included. This may appear more intuitive and easier to understand for individuals not familiar with advanced mathematics or IRT. Note that the base 2 surprisals are used in Equation 9, resulting in the bit unit, but the base can be changed to any other base if one prefers.

One should note that while having a metric scale is in many ways attractive, it also means the scale gets tied to the specific test or questionnaire of a fitted model. This is because the scale is constructed from the IRFs of the items in the test.

7 Test data utilized in the studies

In this section the datasets used throughout the studies in the thesis are presented.

7.1 Swedish SAT

The Swedish SAT, introduced in 1977, is a traditional paper-and-pencil exam that prospective university students can take as part of their application process. It remains a crucial element of the Swedish education system, allowing students to apply to university programs using either their high school grades or Swedish SAT scores. By taking this test, students can compete for university spots in two separate selection groups. The Swedish SAT is administered twice annually, with scores remaining valid for five years. There is no limit on the number of attempts a student can make, and only the highest score is considered for admissions (Lyrén and Hambleton, 2011).

Developing a new version of the Swedish SAT takes about two years. The test comprises two sections: one quantitative and one verbal, each with 80 multiple-choice items. These sections assess general skills such as reading comprehension and quantitative reasoning, which are indicative of potential university success (Lyrén and Hambleton, 2011). All the studies in this thesis utilize data from various administrations of the Swedish SAT.

7.2 National mathematics test

In Sweden, high-school students enrolled in the mathematics 3c course are required to take a national mathematics test. This course is compulsory for students in the natural science and technology programs, but is also available as an elective for students from other programs. Administered at the end of the course, the test significantly influences the final grade. It includes a variety of item types, each contributing to the final score with a varying number of points.

The mathematics tests are constructed at the Department of Applied Educational Science at Umeå University. Every semester, the high-school teachers administrating the tests are asked to report the test results of students born on randomized dates (Skolverket, 2018). Samples from the 2018 and 2019 test administrations are used in the first, second, and fourth study of this thesis.

8 Summary of Papers

8.1 Paper I

Kernel equating (von Davier et al., 2004; Wiberg et al., 2025) is a flexible and powerful technique for equating scores from different test forms. This paper provides an in-depth comparative analysis of kernel equating methods, specifically focusing on the use of log-linear models for presmoothing versus IRT models for presmoothing (Andersson and Wiberg, 2017) when equating mixed-format tests. The study begins with an overview of kernel equating and then explains the methodological differences between log-linear presmoothing and IRT-based presmoothing.

Through simulations and real-data applications using data from both the Swedish SAT and the Swedish national test in mathematics, the paper evaluates the performance of both approaches in terms of accuracy, bias, and robustness. The results demonstrate that the standard errors of the resulting equating functions are smaller for high- and low-performing test takers when using IRT for presmoothing. Simulations also reveal that bias can be smaller or larger for either method depending on what the true equating function is. However, no true equating transformation is known in practical settings, and both presmoothing methods tend to produce reasonable equated curves. Overall, these findings suggest that using IRT models for presmoothing is a viable alternative to log-linear models.

8.2 Paper II

This study investigates the performance of two IRT models for tests with polytomously scored items: the OS model (Ramsay et al., 2020), a nonparametric model rooted in information theory, and the GPC model, a widely used parametric alternative. Using both simulated and real test data, the analysis shows that the OS model consistently demonstrates superior fit across all real datasets (two from the Swedish SAT and two from the national mathematics test). In simulations, the OS model outperforms the GPC model in terms

of bias, but produces larger standard errors for the estimated IRFs. Furthermore, the study examines the application of surprisal arc length, as presented in Section 6.1, to align scores from different IRT models on a common scale. It also demonstrates that surprisal arc length can be a viable alternative to sum scores for scoring test takers' performance.

8.3 Paper III

This study introduces the MMC model, a novel IRT model specifically designed for multiple choice data. The study utilizes both simulated data and real data from the Swedish SAT to empirically demonstrate that the MMC model provides a better fit than the traditional NR IRT model. The paper also illustrates how autoencoders can be used to fit IRT models without making any assumptions about the latent variable scale distributions. Additionally, a modification of surprisal arc length from paper II is proposed, introducing what we refer to as bit scales. Like surprisal arc length, bit scales facilitate score interpretation and comparison across different IRT models on a common metric scale. However, bit scale scores are simpler to understand and also easier to apply when dealing with models for which IRFs tend towards zero or one, for which surprisal arc length tends to infinity. Bit scales are particularly advantageous for models with minimal or no assumptions about the latent variable scale distribution, such as those fitted using autoencoders in this study.

8.4 Paper IV

The fourth study introduces **IRTorch** (Wallmark, 2024), a Python package developed for IRT analysis. This study provides a detailed overview of the package's functionalities, emphasizing its support for a range of IRT models, and illustrates its practical utility through multiple case studies. **IRTorch** is designed to make IRT more accessible beyond the R ecosystem, while also integrating several novel features, such as the MMC model and bit scales discussed in

Figure 5: **IRTorch** logo.

paper III. The package is also built to be extendable, allowing users to develop their own IRT models, estimation algorithms, and latent variable scale transformations.

IRTorch utilizes **PyTorch** (Paszke et al., 2019) for parameter optimization and GPU support. The source code is available on GitHub [1] and the package can be installed through the Python Package Index (PyPI)[2]. An online package documentation website is also available[3]. The package logo is shown in Figure 5.

9 Final Remarks and Further Research

It is not surprising that a more flexible model fits the data better than a parametric model, and one of the primary limitations of studies II and III is that no comparisons were made to other more flexible IRT model alternatives. Examples include kernel smoothing models (Ramsay, 1991) and monotone polynomial models (Falk and Cai, 2016). Another limitation of the second study is that differences in model assumptions are not really considered. It is not completely fair to compare model fit of the OS model, for which monotonicity is not assumed, to other models for which it is. As discussed in Section 3.1, nonmonotone IRFs can make it hard to interpret the latent variable scores, as multiple latent dimensions are likely incorporated into one scale. The practical utility of such models is therefore limited. This is also the reason for similar models not being included in **IRTorch**.

The studies included in the thesis open up several avenues for future research. The MMC model could be easily extended to

[1] `https://github.com/joakimwallmark/irtorch`
[2] `https://pypi.org/project/irtorch/`
[3] `https://irtorch.readthedocs.io/en/latest/`

handle multidimensional latent variables. Multidimensional IRT models could also be utilized to presmooth the score distributions when equating test forms and the unidimensionality assumption does not hold. Additionally, more flexible IRT models such as the MMC could be considered to be used in computerized adaptive or multistage tests (Yan et al., 2014), for which items are selected sequentially in order to maximize information about the latent variable of a test taker using as few items as possible. There are also plans to extend **IRTorch** with more fitting algorithms and models, and potentially also developing a user-friendly web interface for the package to make it even more accessible to researchers and practitioners. Others are also encouraged to suggest improvements or implement changes by forking the public GitHub repository at `https://github.com/joakimwallmark/irtorch` and submitting a pull request.

In summary, the thesis presents a comprehensive exploration of IRT models and their applications, focusing on test equating, model comparisons, and the development of new models.

References

Andersson, B. and M. Wiberg (2017). Item response theory observed-score kernel equating. *Psychometrika 82*(1), 48–66.

Baker, F. B. and S.-H. Kim (2004). *Item response theory: Parameter estimation techniques* (2 ed.). Boca Raton, FL: CRC Press.

Birenbaum, M. and R. A. Feldman (1998). Relationships between learning patterns and attitudes towards two assessment formats. *Educational Research 40*(1), 90–98.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord and M. R. Novick (Eds.), *Statistical theories of mental test scores*, pp. 397–479. Reading, MA: Addison-Wesley.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika 37*(1), 29–51.

Bock, R. D. and M. Aitkin (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika 46*(4), 443–459.

Curi, M., G. A. Converse, J. Hajewski, and S. Oliveira (2019). Interpretable variational autoencoders for cognitive models. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8.

Falk, C. F. and L. Cai (2016). Maximum marginal likelihood estimation of a monotonic polynomial generalized partial credit model with applications to multiple group analysis. *Psychometrika 81*(2), 434–460.

Hambleton, R. K. and H. Swaminathan (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer-Nijhoff Publishing.

Kim, S., M. E. Walker, and F. McHale (2010). Comparisons among designs for equating mixed-format tests in large-scale assessments. *Journal of Educational Measurement 47*(1), 36–53.

Kolen, M. J. and W.-C. Lee (2018). *Mixed-format tests: Psychometric properties with a primary focus on equating (Volume 5)*. CASMA Monograph No. 2.5. Iowa: The University of Iowa.

Kramer, M. A. (1991). Nonlinear principal component analysis using autoassociative neural networks. *AIChE Journal 37*(2), 233–243.

Kullback, S. (1978). *Information theory and statistics*. Gloucester, MA: Dover Publications.

Levy, R. and R. J. Mislevy (2017). *Bayesian psychometric modeling*. Boca Raton, FL: CRC Press.

Lord, F. M. (1975). The 'ability' scale in item characteristic curve theory. *Psychometrika 40*(2), 205–217.

Lord, F. M. and M. R. Novick (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Lyrén, P.-E. and R. K. Hambleton (2011). Consequences of violated equating assumptions under the equivalent groups design. *International Journal of Testing 11*(4), 308–323.

Martinková, P. and A. Hladká (2023). *Computational aspects of psychometric methods: With R* (1 ed.). Boca Raton, FL: CRC Press.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement 16*(2), 159–176.

Paszke, A., S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, and L. Antiga (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems 32*.

Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika 56*(4), 611–630.

Ramsay, J. O. and J. Li (2024). Testgardener: Optimal analysis of test and rating scale data. `https://cran.r-project.org/package=TestGardener`. Version 3.0.0.

Ramsay, J. O., J. Li, and M. Wiberg (2020). Better rating scale scores with information-based psychometrics. *Psych 2*(4), 347–369.

Ramsay, J. O. and B. W. Silverman (2005). *Functional data analysis* (2 ed.). Springer Series in Statistics. New York: Springer.

Rasch, G. (1960). *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests*. Oxford, England: Nielsen & Lydiche.

Reddi, S. J., S. Kale, and S. Kumar (2019). On the convergence of Adam and beyond. *arXiv preprint arXiv:1904.09237*.

Scouller, K. (1998). The influence of assessment method on students' learning approaches: Multiple choice question examination versus assignment essay. *Higher Education 35*(4), 453–472.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal 27*(3), 379–423.

Sijtsma, K. and L. A. van der Ark (2020). *Measurement models for psychological attributes*. CRC Press.

Skolverket (2018). Bedömningsanvisningar, Matematik 3C, kursprov. [Grading Guidelines, Mathematics 3C, Course Exam].

van der Linden, W. J. (2016). *Handbook of item response theory, Volume 1: Models*, Volume 1. CRC Press.

van der Vaart, A. W. (2000). *Asymptotic statistics*, Volume 3. Cambridge University Press.

von Davier, A. A., P. W. Holland, and D. T. Thayer (2004). *The kernel method of test equating*. Statistics for Social Science and Public Policy. New York: Springer.

Wallmark, J. (2024). IRTorch: Item response theory with Python. `https://github.com/joakimwallmark/irtorch`. Version 0.4.2.

Wiberg, M., J. González, and A. A. von Davier (2025). *Generalized kernel equating with applications in R*. Boca Raton, FL: CRC Press.

Yan, D., A. A. von Davier, and C. Lewis (2014). *Computerized multistage testing: Theory and applications*. Boca Raton, FL: CRC Press.