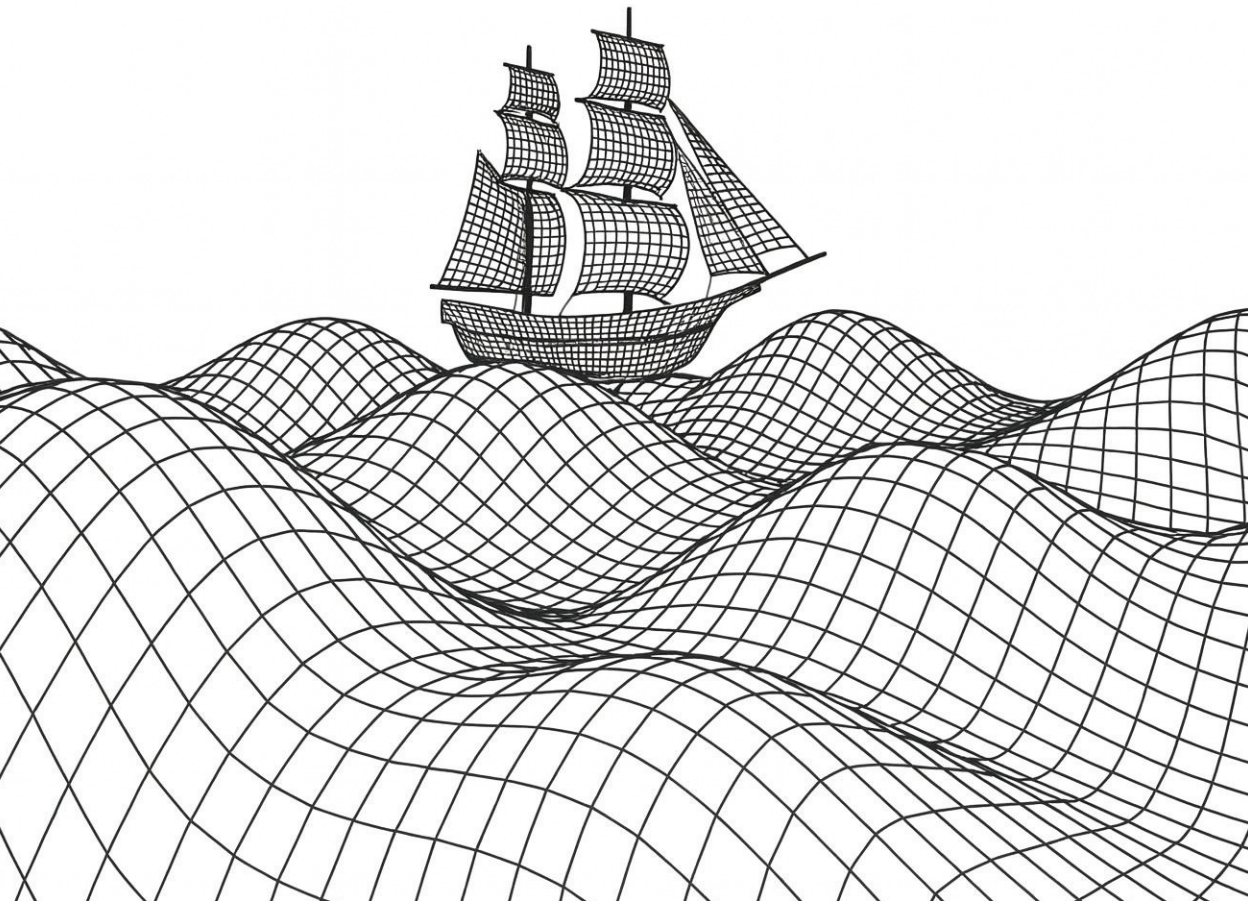


Latent state estimation with longitudinal and adaptive measurements

Karl Sigfrid



Latent state estimation with longitudinal and adaptive measurements

Karl Sigfrid

Academic dissertation for the Degree of Doctor of Philosophy in Statistics at Stockholm University to be publicly defended on Friday 6 March 2026 at 10.00 in lärosal 32, hus 4, Campus Albano, Albanovägen 12.

Abstract

Latent variables are useful constructs for modeling states that cannot be directly observed, such as skills, attitudes and health states. Tests designed to measure latent scores often take the form of questionnaires. The developer of such a test aims to assemble a set of items that measures a latent state with good precision. A challenge is that the best set of items depends on the respondent's true latent score, so a test optimized to give good precision for some respondents may be less precise for others. This dissertation explores statistical methods to optimize the measuring instrument through simulation studies and empirical applications to diverse populations.

The papers included here evaluate adaptive methods that select items based on current knowledge about the respondent. Paper 1 proposes an adaptive method for selecting one item at a time in a Voting Advice Application. With a test that continuously selects the most informative next item, the respondent can conclude the session without answering all items and still get a result that is sufficiently accurate. The proposed method relies on Item Response Theory and a multidimensional latent construct.

In Paper 2, we explored an adaptive model to measure the health states of patients evaluated for symptoms of Parkinson's disease. We compared this adaptive model to optimized static item sets designed for good population-average precision. The Parkinson's dataset consisted of repeated measurements across multiple timepoints, which required a longitudinal approach. In both Papers 1 and 2, the purpose of the methods was to enable more time-efficient versions of tests to increase usage.

Papers 3 and 4 evaluate methods for tracking abilities that change over time. Unlike the Parkinson's scenario with a full test repeated at multiple timepoints, here we have only one observation per time point. In these settings, it is common to abandon traditional statistical models and instead rely on computationally inexpensive algorithms. Of these algorithms, the Elo rating system stands out as the most prominent. This rating system, developed to rate chess players, now has widespread use in many competitive sports and also in education.

We identified limitations associated with the Elo method, and proposed extensions to remedy these. In Paper 3, we developed a hybrid approach that combines standard Elo with statistical modeling to incorporate group-level information. In Paper 4, we demonstrated that in a closed system in which students improve in ability, and where item difficulties are estimated in real time, the Elo method produces increasingly deflated ability estimates. We proposed a method to quantify and offset this system-level deflation.

Keywords: *Longitudinal latent models, Ability tracking, Dynamic ability growth, Elo algorithm, Growth model, MDS-UPDRS, Parkinson's disease, Item selection, Test efficiency, Adaptive testing.*

Stockholm 2026
<http://urn.kb.se/resolve?urn=urn:nbn:se:su:diva-249769>

ISBN 978-91-8107-482-6
ISBN 978-91-8107-483-3



Department of Statistics

Stockholm University, 106 91 Stockholm

LATENT STATE ESTIMATION WITH LONGITUDINAL AND
ADAPTIVE MEASUREMENTS

Karl Sigfrid



Latent state estimation with longitudinal and adaptive measurements

Karl Sigfrid

©Karl Sigfrid, Stockholm University 2026

ISBN print 978-91-8107-482-6

ISBN PDF 978-91-8107-483-3

Printed in Sweden by Universitetservice US-AB, Stockholm 2025

Abstract

Latent variables are useful constructs for modeling states that cannot be directly observed, such as skills, attitudes and health states. Tests designed to measure latent scores often take the form of questionnaires. The developer of such a test aims to assemble a set of items that measures a latent state with good precision. A challenge is that the best set of items depends on the respondent's true latent score, so a test optimized to give good precision for some respondents may be less precise for others. This dissertation explores statistical methods to optimize the measuring instrument through simulation studies and empirical applications to diverse populations.

The papers included here evaluate adaptive methods that select items based on current knowledge about the respondent. Paper 1 proposes an adaptive method for selecting one item at a time in a Voting Advice Application. With a test that continuously selects the most informative next item, the respondent can conclude the session without answering all items and still get a result that is sufficiently accurate. The proposed method relies on Item Response Theory and a multidimensional latent construct.

In Paper 2, we explored an adaptive model to measure the health states of patients evaluated for symptoms of Parkinson's disease. We compared this adaptive model to optimized static item sets designed for good population-average precision. The Parkinson's dataset consisted of repeated measurements across multiple timepoints, which required a longitudinal approach. In both Papers 1 and 2, the purpose of the methods was to enable more time-efficient versions of tests to increase usage.

Papers 3 and 4 evaluate methods for tracking abilities that change over time. Unlike the Parkinson's scenario with a full test repeated at multiple timepoints, here we have only one observation per time point. In these settings, it is common to abandon traditional statistical models and instead rely on computationally inexpensive algorithms. Of these algorithms, the Elo rating system stands out as the most prominent. This rating system, developed to rate chess players, now has widespread use in many competitive sports and also in education.

We identified limitations associated with the Elo method, and proposed extensions to remedy these. In Paper 3, we developed a hybrid approach that combines standard Elo with statistical modeling to incorporate group-level information. In Paper 4, we demonstrated that in a closed system in which students improve in ability, and where item difficulties are estimated in real time, the Elo method produces increasingly deflated ability estimates. We proposed a method to quantify and offset this system-level deflation.

Keywords—Longitudinal latent models, Ability tracking, Dynamic ability growth, Elo algorithm, Growth model, MDS-UPDRS, Parkinson's disease, Item selection, Test efficiency, Adaptive testing

List of Papers

The following papers are included in this thesis.

- PAPER I:** Karl Sigfrid, IRT for voting advice applications: a multi-dimensional test that is adaptive and interpretable. *Quality & Quantity*, 58(5): 4137–4156 (2024).
<https://doi.org/10.1007/s11135-024-01845-6>
- PAPER II:** Karl Sigfrid, Ellinor Fackle-Fornius, Frank Miller, Optimized questionnaire item selection for tracking the progression of motor symptoms in Parkinson’s disease.
(Manuscript)
- PAPER III:** Karl Sigfrid, Ellinor Fackle-Fornius, Frank Miller, Estimating Abilities with an Elo-Informed Growth Model.
(Submitted)
- PAPER IV:** Karl Sigfrid, Ellinor Fackle-Fornius, Frank Miller, Elo estimated abilities with unknown difficulties.
(Manuscript)
-

Acknowledgements

I want to thank my main supervisor, Frank Miller, for his support and guidance throughout the work that resulted in this thesis. He has helped me identify and focus on the essential research problems, and find the direction needed to keep the projects moving forward.

I also want to thank my supervisor, Ellinor Fackle-Fornius, who has helped improve the structure of the papers, and has shared her expertise in latent variable modeling.

It has been a great pleasure to work at the Department of Statistics at Stockholm University, with colleagues who create a friendly atmosphere that makes going to the office a joy. This is a place where people show great enthusiasm for their research and a commitment to giving students a solid understanding of statistical thinking.

Finally, I am grateful to my parents for being enthusiastic and supportive despite having only the vaguest idea about what I am doing.

Contents

Abstract	i
List of Papers	iii
Acknowledgements	v
1 Latent variable analysis	1
1.1 Factor models	2
2 Item Response Theory	3
2.1 The Rasch model	3
2.2 The 2PL, 3PL and 4PL models	4
2.3 The Item Characteristic Curve	4
2.4 Item information	5
2.5 The Graded Response Model	6
2.6 Item Response Theory model estimation	8
2.6.1 Ability estimation with known item parameters	8
2.6.2 Parameter estimation with unknown ability parameters and unknown item parameters	9
2.7 Item Response Theory for adaptive tests	9
3 Longitudinal latent variable models	10
3.1 Extending Item Response Theory models	11
3.2 Learning Factors Analysis and Performance Factor Analysis	11
3.3 Generalized Linear Mixed Models	12
3.4 The Elo model	13
4 Papers	14
4.1 Paper 1: IRT for voting advice applications: a multi-dimensional test that is adaptive and interpretable	15
4.2 Paper 2: Optimized questionnaire item selection for tracking the progression of motor symptoms in Parkinson's disease	18
4.3 Paper 3: Estimating Abilities with an Elo-Informed Growth Model	21
4.4 Paper 4: Elo estimated abilities with unknown difficulties	24
5 Limitations and Future Research	27
6 Sammanfattning	29

1 Latent variable analysis

A latent variable represents an attribute that explains observations, but that cannot be directly observed itself. This type of construct is commonly used in psychometrics to represent traits such as personalities, abilities or attitudes. For instance, the degree of extroversion may help explain a person's social behavior, an ability level may explain the results of an educational test, and an attitude can contribute to a voting behavior. The concept of latent traits can be traced back to the work by Spearman (1904) on general intelligence as a factor behind intellectual achievement (Borsboom et al., 2003). If we measure a latent variable that changes over time, we may refer to the state at one time point as a latent state.

In this dissertation, latent variables are used to describe academic abilities as well as political attitudes, which are both examples of psychometric constructs. Although psychological traits are a common use case for latent variable models, the models are not limited to psychometric measurements. We can also use the value of a latent variable to represent, for example, the severity of a disease state estimated from observations of a patient's disabilities. The latent variable can also represent properties that are not associated with an individual, as when a latent construct is established to represent the quality of a school or the effectiveness of a government.

The concept of a latent variable is distinct from the concept of an index score or other types of summary assessments. A democracy index that scores countries based on voting rights, free speech and rule of law does not necessarily treat democracy as a latent variable. To justify the treatment of democracy as a latent variable, we need to assume that a country has some quality that is not directly observable, but that nevertheless is real and that makes it either more or less likely to have implemented voting rights, free speech and rule of law. The latent trait is thus not merely a summary score of the attributes that we observe. Even though it is possible to interpret a latent trait in terms of the observable variables that it explains, the latent trait represents a common variance of the observations rather than just a weighted sum of them. We can understand that the concept of a latent trait differs from the concept of a weighted sum by imagining two latent trait models that use the same observed variables. If the models are constructed by different researchers, the underlying theory and the interpretations of the constructs may be different between the models even though the models are mathematically equivalent. The researcher who defines a latent variable model may view the latent variable as representing a real phenomenon or as merely a theoretical construct that captures the common variance of the observed variables. The latter approach may be more common in exploratory methods that examine the data without the guidance of a theory.

Bollen and Lennox (1991) distinguish between effect indicator measurement models and causal indicator measurement models. An effect indicator measurement model assumes that the latent trait determines the values of the indicators, as when a greater mathematical ability leads to a higher probability of a correct solution to a math problem. As a contrast to this, Bollen and Lennox (1991) exemplify a causal indicator measurement model with the latent variable socioeconomic status (SES). SES can be associated with indicators such as education, income and residential neighborhood. We view socioeconomic status as a summary measure of the indicators. We do not view SES as an unobservable trait that causes a certain

educational level, income or residential neighborhood.

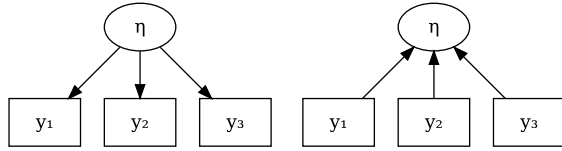


Figure 1: Left: An effect indicator measurement model where the latent variable η determines the indicators. Right: A causal indicator measurement model where η is determined by the indicators.

The latent variable models that we will describe here are for the most part effect indicator measurement models, which assume that the observed indicators are a result of the latent state. In Paper 2, where we use a latent variable to represent the severity of Parkinson’s disease in a patient, it could be argued that the relationship between the latent variable and the indicators fits the description of either an effect measurement model or a causal indicator measurement model depending on whether we interpret the latent state as a measure of an underlying condition or only as a measure of the symptoms.

We need to be aware that there is often a subjective element in the labeling of a latent variable. A latent state that corresponds to a set of indicators can be given different interpretations depending on the theory. If we, for instance, associate a latent variable with indicators that represent responses to mathematical problems, it may be interpreted as a measure of knowledge in the field of mathematics, or it may be interpreted as a measure of general logical ability. A latent state associated with indicators of economic growth may be interpreted as a measure of an effective economic policy or as a measure of entrepreneurial culture. Thus, when latent variable models are used to explain observations, we can rarely rule out alternative explanations.

An assumption of latent variable models is local independence. Given the value of the latent score, the observed variables are independent (Borsboom et al., 2003). Whereas the probability of a correct response to two different test questions both depends on the respondent’s ability, a correct response to one of the items provides no additional information about the probability of responding correctly to the other item.

1.1 Factor models

The factor model, used in Confirmatory Factor Analysis (CFA), is a basic example of a latent variable model. The factor model illustrated in Figure 2 can be formalized with three equations.

$$\begin{aligned} y_1 &= \alpha_1 + \lambda_1 \eta + \delta_1, \\ y_2 &= \alpha_2 + \lambda_2 \eta + \delta_2, \\ y_3 &= \alpha_3 + \lambda_3 \eta + \delta_3, \end{aligned} \tag{1}$$

where α_i is an intercept, λ_i a loading, and δ_i the residual for indicator y_i . The residual δ_i represents the variance in the indicator not accounted for by the latent

variable η . If both the latent variable and the indicators are standardized, λ_i is interpreted as the correlation between the latent variable and indicator y_i . The square of λ_i is then the proportion of the variance in the indicator explained by the latent variable.

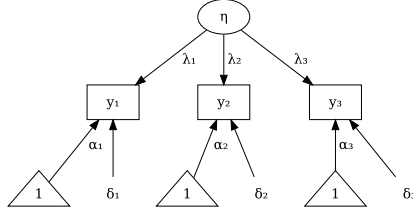


Figure 2: An illustration of a confirmatory factor model. The indicators y_1 , y_2 , y_3 are assumed to depend on the latent variable η , which represents the indicators' shared variance. The parameter λ_i is a factor loading. Each indicator has a residual variance δ_i which is not explained by η . Each indicator also has an intercept α_i

2 Item Response Theory

Related to confirmatory factor analysis is Item Response Theory, used to estimate latent traits based on ordinal data. In an educational setting, the latent trait measured in an IRT model is typically an ability, such as a mathematical ability. The data is often gathered through the administration of a test, where each test question is referred to as an item. However, Item Response Theory is useful also in numerous other contexts outside of education. In the first paper of this dissertation, we use an Item Response Theory model to place a respondent on an ideological scale. The location of the respondent on the ideological scale is then estimated based on responses that indicate an attitude to political statements relevant to the voting choice.

In standard univariate IRT models, we assume that the items in a test measure the same latent trait. We also assume independence between observations conditioned on the model parameters.

2.1 The Rasch model

The simplest IRT model is the Rasch model. If we have a respondent j and a question item i , the Rasch model is defined by the equation

$$P(y_{i,j} = 1) = \frac{e^{\theta_j - d_i}}{1 + e^{\theta_j - d_i}}, \quad (2)$$

where θ_j is the ability of the respondent and d_i the difficulty of the item. The variable $y_{i,j}$ is 1 if the response is correct and otherwise 0. Thus, $P(y_{i,j} = 1)$ is the probability of a correct answer. This probability depends on the item difficulty and the ability of the respondent, which are both measured on the same scale. When the respondent ability equals the item difficulty, the probability of a correct answer is 0.5.

2.2 The 2PL, 3PL and 4PL models

Whereas the Rasch model acknowledges that each item has its own difficulty with respect to the measured latent trait, it does not consider that different items can have different discrimination abilities. A higher discrimination ability implies that the item is better at distinguishing higher-ability respondents from lower-ability respondents in the region of the item difficulty. We move from a Rasch model to a 2-parameter logistic (2PL) model by including the discrimination parameter a_i . The model then becomes

$$P(y_{i,j} = 1) = \frac{e^{a_i(\theta_j - d_i)}}{1 + e^{a_i(\theta_j - d_i)}}. \quad (3)$$

If θ_j is known, The 2PL model is equivalent to the standard logistic model. Sometimes a third item parameter g_i is added to the model, which then becomes the 3PL model, with the equation

$$P(y_{i,j} = 1) = g_i + (1 - g_i) \frac{e^{a_i(\theta_j - d_i)}}{1 + e^{a_i(\theta_j - d_i)}}. \quad (4)$$

The parameter g_i , commonly called the guessing parameter, represents the lower asymptote of the probability of a correct answer when for very low ability levels. It accounts for the possibility that a respondent with very low ability selects the correct answer by random chance or by making a qualified guess based on partial knowledge. A fourth parameter can likewise be added in the 4PL model to represent an upper asymptote of the probability that a respondent with very high ability responds correctly. In the equation

$$P(y_{i,j} = 1) = g_i + (u_i - g_i) \frac{e^{a_i(\theta_j - d_i)}}{1 + e^{a_i(\theta_j - d_i)}}, \quad (5)$$

the parameter u_i is the asymptotic probability of a correct answer for very high ability levels. This asymptotic probability is less than 1 since a highly skilled respondent may be careless or affected by conditions like stress and fatigue.

The parameter terminology is heavily influenced by educational applications, but can often be well understood in other contexts such as the measurement of attitudes. If we measure a respondent's opinions on a political scale, the items may be statements of a political nature, and the response options indicate to what degree the respondent agrees. With a latent trait that measures conservatism, holding more conservative views is analogous to a higher ability, and holding less conservative views is analogous to a lower ability. A very difficult item is then an item that only very conservative respondents will agree with, and an easy item is an item that also non-conservative respondents can agree with. An item with a high discrimination parameter value is an item that serves well to distinguish more conservative from less conservative respondents. An item with a low discrimination parameter value discriminates less well, which is an indication that the item may not measure conservative beliefs.

2.3 The Item Characteristic Curve

The probability of answering an item correctly is a function of the respondent's ability θ_j . The function is monotonically increasing with θ_j , i.e., the probability

of a correct answer will never decrease with greater ability. The curve that illustrates how the probability of a correct answer changes with θ is called the Item Characteristic Curve (ICC).

As illustrated in Figure 3, the ICC is assumed to have a logistic shape. This shape implies that the probability changes more rapidly with θ when θ is close to the item difficulty d_i , and the change in probability flattens out as the distance between ability and difficulty increases in either direction. The logistic shape makes intuitive sense, as it will make little difference if an already overqualified respondent gains additional skills. Likewise, an underqualified respondent who already has a low chance of success does not have much success probability to lose from a further decrease in skill level.

In the 2PL model, the ICC is defined by the parameters that represent item difficulty d_i and discrimination a_i . Two items with the same value for the a_i while different with regard to d_i will be shifted. As shown in the middle graph in Figure 3, the ICC of an easier item is shifted to the left, and the ICC of a more difficult item is shifted to the right. If we instead have two items with the same difficulty, but different values for the discrimination parameter a_i , the ICCs will intersect at their common difficulty level on the x-axis and at 0.5 on the y-axis. This scenario is illustrated for three items in the right chart in Figure 3.

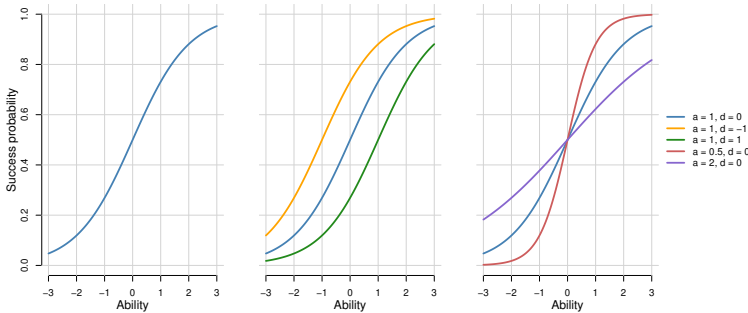


Figure 3: Left: A reference Item Characteristic Curve for an item with difficulty 0 and discrimination 1. Middle: The reference ICC together with one less difficult item and one more difficult item. The three items have discrimination parameters with the same value. Right: The reference ICC, one item with lower discrimination and one item with higher discrimination. The three items have the same difficulty.

2.4 Item information

In IRT, a common end goal is to estimate the abilities of respondents with as little uncertainty as possible. The asymptotic variance of the Maximum Likelihood estimate of the ability is the inverse of the Fisher information obtained from all the given responses (Reckase, 2009). Therefore, to minimize the asymptotic variance of the ability estimate, we want to maximize the Fisher information.

The Fisher information of an item i can be expressed as a function that depends on the respondent's true ability. In the 2PL model, the Fisher information obtained

about a respondent j from a response to item i can be calculated as

$$I_i(\theta_j) = a_i^2 P_i(\theta_j) (1 - P_i(\theta_j)), \quad (6)$$

where $P_i(\theta_j) = P(y_{i,j} = 1)$. Figure 4 illustrates the Fisher information functions of three different items. In the figure, we can see that the top of each information curve occurs at the value of d_i . Thus, we obtain the most information about a respondent when the difficulty of the item equals the ability of the respondent.

We should note that θ_j is unknown in most real-world scenarios. The same is true of the item parameters. Therefore, we typically use estimated values in the calculation, which means that we obtain an estimate of the Fisher information, and therefore of the variance of the latent trait estimate.

We are often interested in the precision of the ability estimate of a respondent who has answered multiple items. The Fisher information obtained from the responses to a set of items is the sum of the Fisher information from each item in the set. It is calculated as

$$I_{i \in U_i}(\theta_j) = \sum_{i \in U_i} I_i(\theta_j), \quad (7)$$

where U_i is the set of items that have been answered.

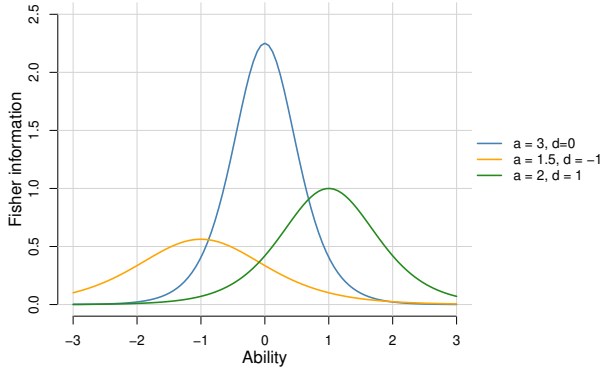


Figure 4: The Fisher information with regard to an item obtained from an item response is a function of the true ability. The information function is different with different item parameters.

2.5 The Graded Response Model

In the 2PL model, the probability of a correct answer to an item is a function of the respondent ability, the item difficulty and the item discrimination parameter, as formulated in Equation 3. This assumes that only two outcomes are possible: success and failure. Whereas this is sufficient for many applications, we sometimes need to consider a greater number of outcomes. For instance, an item on a math test may give at most 4 points, and depending on how good the solution is, it is possible to receive any score from 0 to 4. We can model this as an ordinal scale with 5 levels. An ordinal scale with 5 levels can also be described as an ordinal scale

with 4 thresholds as illustrated in Figure 5. A full score is equivalent to passing all thresholds. A score one less than the full score is equivalent to passing all except the most difficult threshold. A score of one means that the respondent has passed only the easiest threshold, and a score of zero means that none of the thresholds were passed. The 2PL model is a special case of this, in which the ordinal scale has only two levels and one threshold.

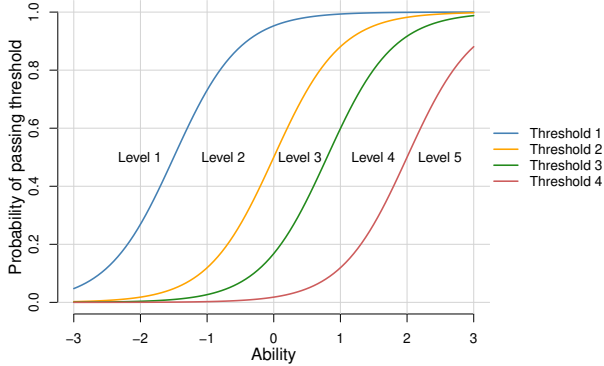


Figure 5: The 4 thresholds of a Graded Response Model with 5 levels. Each probability curve represents the probability of passing a threshold; that is, the curve associated with threshold 1 represents the probability of reaching level 2 or higher.

With an ordinal response scale that has 5 levels, we do not model the probability of success but rather 5 different probabilities, one for each possible outcome. This is done in two steps. In step one, we calculate, for each threshold, the probability of passing this threshold with the formula

$$P(y_{i,j} \geq m) = \frac{e^{a_i(\theta_j - d_{i,m})}}{1 + e^{a_i(\theta_j - d_{i,m})}}. \quad (8)$$

This is the same formula as the 2PL formula, with two important exceptions. The probability denoted on the left-hand side is here the probability of reaching level m or higher on the ordinal scale. We have also replaced d_i with $d_{i,m}$ as each level is associated with its own difficulty. In contrast, we still have only one parameter value a_i for the ordinal item, implying that the discrimination parameter value is constrained to be the same for all thresholds. We calculate the probability of a specific response option m as

$$P(y_{i,j} = m) = P(y_{i,j} \geq m) - P(y_{i,j} \geq m + 1), \quad (9)$$

which is the probability of reaching level m or higher minus the probability of reaching level $m + 1$ or higher. The interpretation of this is that a response at level m demonstrates enough skill to reach any level up to and including m , but not enough skill to reach level $m + 1$ or above.

2.6 Item Response Theory model estimation

2.6.1 Ability estimation with known item parameters

If item difficulties are known, we can use a Maximum Likelihood procedure to estimate the ability of a respondent. With the 2PL model, we use Formula 3 as the starting point to derive the likelihood of all the responses of a test taker. Under the standard IRT assumption that all responses are independent conditioned on the ability, we can formulate the likelihood for the ability of test taker j as

$$L(\theta_j|\mathbf{y}, \mathbf{a}, \mathbf{d}) = \prod_{i=1}^I P(y_{i,j}), \quad (10)$$

where \mathbf{y} represents all responses, and \mathbf{a}, \mathbf{d} the parameters of all items answered by the test taker. I is the number of items that respondent j has answered. In more explicit form, the likelihood can be written

$$L(\theta_j|\mathbf{y}, \mathbf{a}, \mathbf{d}) = \prod_{i=1}^I \frac{e^{a_i(\theta_j - d_i)^{y_{i,j}}}}{1 + e^{a_i(\theta_j - d_i)}}. \quad (11)$$

Note that in this equation, $y_{i,j}$ in the numerator is 1 for correct answers and 0 for incorrect answers. As a result, the full expression will be the probability of the outcome, whether successful or not. If all item parameters are known, finding the Maximum Likelihood (ML) estimate of the ability is a simple matter. Likewise, if the abilities of the respondents are known, it is equally simple to estimate the item parameters.

Common alternatives to ML estimates are Maximum a Posteriori (MAP) estimates and Expected a Posteriori (EAP) estimates. Both these estimates are extracted from a Bayesian posterior distribution, i.e., a distribution that combines the likelihood function with a distribution that reflects our prior beliefs about the probability distribution of the estimate. A common reason for using a prior is that the available data is insufficient to get a reliable ability estimate solely from the likelihood function. In the case where a respondent answers all items correctly, or answers all items incorrectly, we cannot obtain a finite ML estimate as the function is maximized by an infinitely large or infinitely small ability.

When we calculate a posterior function rather than just a likelihood function, the prior will reflect what we either know or believe about the respondent's ability prior to the data analysis. If we know nothing about the respondent, we may have useful knowledge about a larger population of respondents. The posterior is calculated

$$f_{\text{post}}(\theta|\mathbf{y}) = \frac{L(\theta|\mathbf{y}) \cdot f_{\text{prior}}(\theta)}{f(\mathbf{y})}. \quad (12)$$

In implementation, you can disregard the denominator $f(\mathbf{y})$, as it has a constant value and does therefore not affect the posterior ability estimate.

When we calculate the MAP estimate, we locate the mode of the posterior ability distribution. Many algorithms exist for finding the maximum of a concave function, and functions are available in any major programming language used for data analysis. An alternative to the MAP estimate is the EAP estimate, which is the expected value of the posterior distribution. In IRT applications, the posterior is typically quite symmetric, and therefore the choice between the MAP estimate and the EAP estimate may be of limited importance.

2.6.2 Parameter estimation with unknown ability parameters and unknown item parameters

Whereas it is easy to estimate respondent abilities when the item parameters are known, the true item parameter values are often not available. Instead, what we have is a series of response patterns, from respondents of unknown ability, who have answered items of unknown difficulty.

Several approaches have been developed to make estimations in this type of cold-start scenario. Examples are Markov Chain Monte Carlo (MCMC) sampling, the Expectation-Maximization (EM) algorithm, and Maximum Marginal Likelihood (MML) estimation (Baker & Kim, 2004).

Regardless of approach, the model is not identifiable unless we constrain either the ability parameters or the item parameters. In Equation 11, we see that the probability of a correct answer is a function of the difference between the ability θ_j and the difficulty d_i . Given a_i , for any combination of θ_j and d_i , we can calculate a probability of a correct answer. If we add any value to θ_j , we will get the same probability if we also add the same value to d_i . Therefore, without any constraints we have an infinite number of parameter combinations that result in the same likelihood.

With the MML approach, we can estimate the parameters in two steps, where we start with the item parameters. As we don't know the abilities of the respondents we initiate the process with the assumption that the abilities in the group of respondents follow a certain distribution, typically a standard normal distribution. This distributional assumption is the constraint that makes the model identifiable. Using Formula 11, we integrate the expression over the assumed ability distribution. Our item parameter estimates are those that maximize the expression.

In Paper 4, where we use the Rasch model, we approximate the integral by adding a scaling parameter to the equation for the probability of a correct answer. We assume that the abilities of the respondents follow a distribution that we refer to as $g(\theta_j)$. The likelihood function

$$L(d_i|\mathbf{y}) = \prod_{j \in J} \int_{-\infty}^{\infty} \frac{(e^{\theta_j - d_i})^{y_{i,j}}}{1 + e^{\theta_j - d_i}} g(\theta_j) d\theta_j, \quad (13)$$

where we integrate the expression over $g(\theta_j)$, can then be approximated by

$$L(d_i|\mathbf{y}) \approx \prod_{j \in J} \frac{(e^{-d_i/s})^{y_{i,j}}}{1 + e^{-d_i/s}}, \quad s = \sqrt{1 + \frac{3}{\pi^2}}. \quad (14)$$

To calculate the scaling factor s , we use a method of moments transformation to approximate the logistic distribution with a normal distribution. This procedure is used in Glickman (1999) to update ability estimates with a closed form expression.

2.7 Item Response Theory for adaptive tests

In a traditional test, the respondent answers a set of items. For every respondent, the items are the same, and they are answered in the same order. An adaptive test, in contrast, does not determine in advance which items to answer and in what order. Instead, it adapts its behavior based on the responses given thus far. Typically

adaptive tests are digital, even though it is possible to implement adaptivity in a pen-and-paper test.

An obvious use case for an adaptive test is a computer-based diagnostic test of a skill, such as mathematics. After each response, the test software updates its estimate of the respondent's ability. It then presents a new item with an appropriate difficulty level, which here means that the item is not so easy or so difficult that we can know beforehand with high certainty whether the respondent will succeed. From a purely statistical perspective, we often prefer an item that we believe gives the respondent a 50 percent chance of success. In an IRT framework, this maximizes the expected information obtained from the answer. In practice, a test where the respondent fails on half the items may be too difficult to maintain motivation. Therefore, test designers sometimes prefer an item difficulty that places the probability of a correct answer in the region of 0.7 (Bolsinova et al., 2026) or 0.75 (Straatemeier, 2014).

At the beginning of the adaptive test, the system has not collected any information about the ability of the respondent. To match the respondent with a first item, the system needs to make an initial guess about the ability, for instance that the respondent's ability equals the average ability in a reference population. After this initial guess, as shown in Figure 6, the respondent answers the selected question, which allows the system to update the ability estimate. Based on this new ability estimate, a second question is selected. The system cycles between selecting new items and updating the ability estimate until the test terminates. The criteria for terminating the test depend on the purpose of that particular test. For example, a test may terminate when the precision of the ability estimate passes a pre-determined threshold.

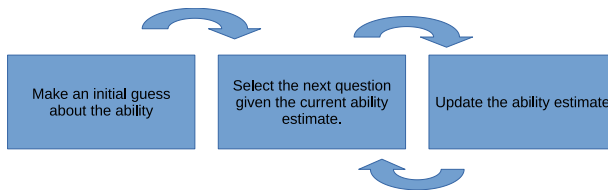


Figure 6: The overall structure of an adaptive test.

3 Longitudinal latent variable models

In a typical test scenario, respondents answer a set of items in one session. Each outcome is then a manifestation of the ability that we want to estimate, so if the test includes 20 items the ability is estimated based on 20 observations. An implicit assumption is that the ability remains constant throughout the test. If this is not the case, and the ability changes during the test session, the classical IRT model no longer makes sense, as the multiple outcomes are no longer associated with one single ability.

In an Intelligent Tutoring System (ITS), a system that teaches a skill and adapts its behavior to the estimated user ability, the user may solve exercises that are embedded between teaching blocks. If learning happens between one exercise item

and the next, then we have only one outcome per ability. Therefore, we need an approach that acknowledges that the ability at each time t is unique while also allowing borrowing of information across time. This is what many types of longitudinal models aim to achieve. In this section, a few such models will be described.

3.1 Extending Item Response Theory models

The basic Rasch model calculates the probability of a correct answer as

$$P(y_{i,j} = 1) = \frac{e^{\theta_j - d_i}}{1 + e^{\theta_j - d_i}}, \quad (15)$$

where θ_j is the ability of the respondent. In this equation, θ_j is treated as a constant value. If we instead want to allow the ability to change over time, we can replace the constant ability with an expression that depends either on time or on other variables that are allowed to differ from one time point to the next.

3.2 Learning Factors Analysis and Performance Factor Analysis

In Learning Factors Analysis (LFA) (Cen et al., 2006), the ability of respondent j is modeled as a linear function of the number of answers that respondent j has given. In the simplest case where we only measure a single ability, or a single knowledge component as it is referred to in LFA, the probability of a correct answer at time t is calculated as

$$P(y_{i,j,t} = 1) = \frac{e^{\alpha_j + \gamma \cdot m_{j,t} - d}}{1 + e^{\alpha_j + \gamma \cdot m_{j,t} - d}}, \quad (16)$$

where we can see that θ_j has been replaced by the expression $\alpha_j + \gamma \cdot m_{j,t}$. The parameter α_j is an individual intercept for respondent j , and the parameter γ is the learning rate, assumed to be the same for all respondents. The variable $m_{j,t}$ is the number of items that respondent j has answered by time t .

In an LFA model, an item may involve multiple knowledge components, i.e., solving the item may require multiple skills that each is associated with its own difficulty. The general form of the LFA equation that incorporates multiple knowledge components can be found in the supplementary materials to Paper 3. In LFA with multiple knowledge components, each knowledge component is associated with one difficulty. In the special case with just one knowledge component, all items will in effect have one and the same item difficulty, represented by d . The assumption that a difficulty is associated with a knowledge component, rather than with an individual item, makes it appropriate to limit the scope of each knowledge component so that items sharing one knowledge component are sufficiently homogeneous.

LFA does not distinguish between correct and incorrect responses. This makes it different from the otherwise similar Performance Factor Analysis (PFA) (Pavlik et al., 2009). In the special case with a single knowledge component, PFA models the probability of a correct answer as

$$P(y_{i,j,t} = 1) = \frac{e^{\gamma \cdot s_{j,t} + \rho \cdot f_{j,t} - d}}{1 + e^{\gamma \cdot s_{j,t} + \rho \cdot f_{j,t} - d}}, \quad (17)$$

where $s_{j,t}$ is the number of correct answers by respondent j up to time t . The parameter γ determines how much the estimated ability increases with each correct answer. The variable $f_{j,t}$ represents the number of incorrect responses up until time t , and ρ the effect of each incorrect response on the ability estimate. As in the LFA equation, d is the difficulty parameter that unidimensional items associated with the same knowledge component have in common. Pavlik et al. (2009) suggest fitting the model parameters with a Maximum Likelihood procedure.

Whereas PFA extracts more information from the data than LFA, as it distinguishes successful attempts from failed attempts, it does not account for different item difficulties within the same knowledge component. An important difference between PFA and LFA is that PFA does not include an individual intercept in the expression for a respondent's ability at time t . This makes the method easier to implement when the goal is to track new respondents with unknown initial abilities.

3.3 Generalized Linear Mixed Models

Generalized Linear Mixed Models (GLMM) are a broad family of models that can be used in a variety of settings. They differ from standard Generalized Linear Models (GLM) in that they incorporate random effects that account for the variation between groups of observations (McCulloch, 2003). As a generalized linear model, it can handle response data that is not normally distributed, such as a Bernoulli distributed variable that represents correct and incorrect answers to an item (McCulloch, 2003). With a Bernoulli distributed response variable, the general form of the GLMM model is

$$P(y_{i,j} = 1) = \frac{e^{\mathbf{x}_i^\top \boldsymbol{\beta} + \mathbf{z}_{i,j}^\top \mathbf{u}_j}}{1 + e^{\mathbf{x}_i^\top \boldsymbol{\beta} + \mathbf{z}_{i,j}^\top \mathbf{u}_j}}, \quad (18)$$

$$\mathbf{u}_j \sim \text{Normal}(\mathbf{0}, \boldsymbol{\Sigma}),$$

where i is the index of the observation and j is the index of a group to which observation i belongs. The vectors $\mathbf{x}_{j,t}$ and $\mathbf{z}_{j,t}$ are known covariates, $\boldsymbol{\beta}$ an unknown parameter vector, \mathbf{u}_j a random effects vector with mean $\mathbf{0}$ and a $p \times p$ variance-covariance matrix $\boldsymbol{\Sigma}$, where p is the number of random effects associated with a respondent.

A typical use case for GLMM models is a national achievement test where a school effect can be modeled as a random effect. The results at different schools are then assumed to vary around the national mean. In our application, where we use GLMM to track motor symptoms over time in patients evaluated for Parkinson's disease, we have measurements from each individual at multiple timepoints. Observations associated with the same individual are grouped together, and individual trajectories are modeled as random effects.

A natural question to ask is why we cannot treat groups of observations within the fixed-effects framework, with a categorical group variable and interaction terms to incorporate group effects of covariates. Whereas this is an option, an advantage of using random effects for prediction is that the distribution of group effects allows us to borrow information across groups. This can be especially valuable when fewer observations per group are available.

We can obtain estimates of the model parameters $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ by maximizing the likelihood

$$L(\boldsymbol{\beta}, \Sigma | \mathbf{y}) = \prod_j \int \prod_i \frac{\exp(\mathbf{x}_i^\top \boldsymbol{\beta} + \mathbf{z}_{i,j}^\top \mathbf{u}_j)^{y_{i,j}}}{1 + \exp(\mathbf{x}_i^\top \boldsymbol{\beta} + \mathbf{z}_{i,j}^\top \mathbf{u}_j)} \times (2\pi)^{-p/2} \det(\Sigma)^{-1/2} \exp(-\mathbf{u}_j^\top \Sigma^{-1} \mathbf{u}_j / 2) d\mathbf{u}_j. \quad (19)$$

The outcome variable $y_{j,t}$ is either 0 or 1. When it is 1, the logistic function gives the probability of a correct answer. When it is 0, the numerator becomes 1, and the logistic function gives the probability of an incorrect answer. The integral in Expression 19 can be approximated numerically.

A longitudinal IRT model can be formulated in GLMM form as

$$P(y_{i,j,t} = 1) = \frac{e^{\mathbf{x}_{i,j,t}^\top \boldsymbol{\beta} + \mathbf{z}_{i,j,t}^\top \mathbf{u}_j - d_{i,t}}}{1 + e^{\mathbf{x}_{i,j,t}^\top \boldsymbol{\beta} + \mathbf{z}_{i,j,t}^\top \mathbf{u}_j - d_{i,t}}}, \quad (20)$$

where the time index t has been added. Compared to Equation 18, we have also added the term $d_{i,t}$ to represent the difficulty of the item encountered at time t . The expression $\mathbf{x}_{i,j,t}^\top \boldsymbol{\beta} + \mathbf{z}_{i,j,t}^\top \mathbf{u}_j$ represents the ability of respondent j at time t . Thus, Equation 20 is a Rasch model, in which the probability of a correct answer is a function of the difference between the ability at time point t and an item difficulty. In this model, the ability has both a fixed component and a random component. GLMM models of this type are used in Papers 2 and 4, with slightly different specifications. Whereas in Paper 3, the GLMM model is used for benchmarking purposes, it is the central model in Paper 2 where we use it to track the progression of motor symptoms in Parkinson's disease patients. The model can be fitted by MCMC sampling. In doing this, we can incorporate Bayesian priors.

In a model where time is the only covariate, the expression $\mathbf{x}_{i,j,t}^\top \boldsymbol{\beta}$ represents the mean ability at time t for all respondents, and $\mathbf{z}_{i,j,t}^\top \mathbf{u}_j$ represents the deviation of individual j from this mean. Depending on how a GLMM model is implemented, it may or may not assume a specific shape of the ability growth curve. If the time variable is a categorical variable, the mean ability at each time point is assigned its own parameter, which allows the growth curve to have any shape. GLMM can therefore be a viable option to track the ability development of a student over time also when we do not want to assume a shape of the growth curve.

3.4 The Elo model

The Elo rating system was originally developed to rate chess players (Elo, 1978). It is still used for this purpose, but is also popular in other settings where we are interested in tracking abilities that change over time, such as computer games and education.

The Elo model is not a typical statistical model. Compared to many other approaches, it has the advantage of being both simpler and less computationally demanding. However, it also has the disadvantage that some of its statistical properties are unknown (Bolsinova et al., 2026). An additional weakness is that estimates of an individual's ability are made in a manner that does not incorporate group-level information.

The Elo method and the Rasch model share the basic equation for the probability of a correct answer. Thus, this probability is calculated as

$$P(y_{i,j} = 1) = \frac{e^{\theta_j - d_i}}{1 + e^{\theta_j - d_i}}. \quad (21)$$

This version of the formula is a slight modification of the formula typically used in chess rating systems. In a chess context, the expression $e^{\theta_j - d_i}$ will instead be $10^{(R_1 - R_2)/400}$, where R_1 and R_2 are the abilities (ratings) of two chess players who compete in a game. It is a simple matter to transform an ability between the Rasch model and the chess version of the equation.

Notably, θ_j corresponds to R_1 , and d_i to R_2 . This means that the ability of a student in an educational setting corresponds to the ability of chess player 1, while the item difficulty corresponds to the ability of chess player 2. We can view the educational application of the Elo method as a chess analogy where the student is a chess player and the item is the opponent player. A correct answer to the item is then analogous to winning the chess game, and an incorrect answer is analogous to a defeat (Klinkenberg et al., 2011). If both the respondent ability and the item difficulty are unknown, both can be simultaneously estimated. The updates are done with the expression

$$\theta_t = \theta_{t-1} + K(y_{t-1} - E(y_{t-1})), \quad (22)$$

where the expected outcome $E(y_{t-1})$ equals the estimated probability of a correct answer. We use t to index time points. The expression implies that a respondent who succeeds when the expectations are low will make a greater gain in estimated ability than if the expectations were high. This makes intuitive sense, as high expectations of a correct answer indicate that we already knew that the respondent was likely to succeed, and had this information incorporated in the estimated ability. With the same logic, a respondent who fails will make a greater loss in estimated ability if the expectations of a correct answer were high than if the expectations were low.

The Elo method plays a central role in Paper 3, where we develop a model that can improve the accuracy of Elo estimates. Paper 4 explores how the Elo method under some circumstances causes abilities to become increasingly underestimated over time.

4 Papers

In many contexts, it is useful to continuously make measurements and update a latent trait estimate based on the outcome. These updated estimates allow practitioners, or computer systems, to be better informed and able to provide appropriate interventions. In an educational setting, these interventions can be instructions and practice exercises at the right difficulty level. In an adaptive test setting, the intervention can be to select the most informative item from an item pool to present. This dynamic approach is an alternative to a static process where each step is determined in advance based on an initial assessment.

The usefulness of a dynamic approach depends on how different the subjects are, and whether a tailored intervention gives enough added value to justify the

extra complexity. The usefulness is also determined by whether different items in the item pools are better suited for different individuals, or whether some items tend to dominate and be better for most users regardless of individual differences.

To be practical, the implementation of a dynamic system often requires digital assessments, so that updated estimates and selections of interventions can be performed automatically.

Both Papers 1 and 2 in this dissertation explore methods for adaptive item selection. In Paper 1, adaptive item selection is used in the context of Voting Advice Applications (VAA), to enable VAAs that require fewer answers while they are still able to rank alternatives with sufficient accuracy (Sigfrid, 2024). Whereas a VAA is a special case of a recommender system that ranks parties or candidates ahead of an election, the proposed method can be applied to a wide variety of recommender systems. In Paper 2, dynamic item selection is employed to select the most informative items from a questionnaire that aims to measure the symptom severity for patients who participate in research on Parkinson’s disease.

In Papers 3 and 4, we propose methods to improve the accuracy when we track abilities that change over time. While both these papers address the same practical problem, they target different subproblems. In Paper 3, we describe a dynamic method to track abilities when item difficulties are known. In Paper 4, we abandon the assumption of known item difficulties and explore on-the-fly estimation of both abilities and item difficulties with the Elo method (Klinkenberg et al., 2011). We identify a major weakness in on-the-fly estimation with standard Elo, and propose a method to mitigate this problem.

4.1 Paper 1: IRT for voting advice applications: a multi-dimensional test that is adaptive and interpretable

A test that aims to match the preferences of a voter to a political party or a candidate is referred to as a voting advice application (VAA). VAAs have become popular to use in election times, as they help voters orient themselves in the political landscape. The adoption is particularly large in northern European countries.

A VAA presents a sequence of politically loaded statements, and the user answers to what extent they agree with each statement. The response scale is often a Likert scale with five response options ranging from “strongly disagree” to “strongly agree”. After completion of the test, the VAA calculates how close the user is to each party or candidate.

To compare the opinions of the respondents to those of the political parties, the opinions of each party must be recorded. To record party positions, a representative from each party may respond to the same set of items as the end user. The party position can also be defined or adjusted by experts to ensure that the stated party positions are consistent with the party policies.

A VAA can match respondents to political parties using a summed similarity score. Proximity can also be measured as the distance on a unidimensional or multidimensional latent trait scale. We will here focus on the latent trait approach, where each latent trait represents the location along a political scale.

Item Response Theory models are well suited to estimate the latent traits measured by a VAA. However, when we move from the domain of education, where IRT is most commonly applied, into the domain of political opinions, we need to reinterpret the IRT model parameters. Whereas in an education setting, the ability

parameter literally measures an ability, in the VAA setting it measures the location on a political scale. For instance, an ability could represent a location on the traditional economic left-right scale where the left prefers more resources to government programs and a more regulated economy, while the right prefers lower taxes and less regulation. Whether to equate higher ability with opinions further to the left or further to the right is an arbitrary choice. In the example below, a higher ability is analogous with views that are further to the right. Using this analogy, a more difficult item is a statement that is more difficult to agree with, i.e., only a respondent located further to the right is likely to agree. A less difficult item is a statement that is easier to agree with, so that also respondents further to the left are likely to agree. An item with high discrimination capability is in this analogy an item that is good at differentiating respondents who are at different locations on the left-right scale.

In any IRT model, scale validation is essential. If we plan to use a set of items to measure an ability, we should make sure that the items actually measure the same construct. Germann et al. (2015) has pointed out that VAAs often fail in this respect by defining political scales based only on prior beliefs, without support in the data. We can validate a scale either with more formal tests or with heuristic methods. In Paper 1, we used two political scales – one that measured left vs right economic sentiments and one that measured socially liberal vs socially conservative sentiments. To validate these scales, we used Mokken scale analysis (MSA) (Mokken, 1971). A widely used tool in MSA is the scalability coefficient, also referred to as the Loevinger's coefficient or the coefficient H (Loevinger, 1948; Sijtsma and van der Ark, 2017; Watson et al., 2012). The coefficient H measures how well a group of items work together as a unidimensional scale. Rules of thumb are used to determine whether the justification for using the item set to measure a latent trait is weak, moderate or strong. To confirm that a single item belongs to the scale, we can use a more heuristic summed score technique: We divide respondents into groups based on their overall summed score, i.e., the summed score from all items except the item of interest, then verify that groups of respondents with higher overall scores on average perform better with the item of interest.

VAAs often require respondents to answer a large set of items. This works well for some respondents, while others find the response burden too heavy. Survey fatigue, which is the exhaustion from filling out long questionnaires, is a phenomenon that has been studied, and it disproportionately affects individuals with lower education and who are not native speakers of the language used in the survey.

To be more accessible, a VAA could allow respondents to conclude the test without answering all items. This would lessen the response burden, but at the price of matching results that are based on less data. With fewer responses, it is more important that those responses are as informative as possible. A reduced item set, if a subset of the original set, will always result in a less precise result. However, a good method for selecting those items can reduce the loss in precision.

Item Response Theory provides a simple way of determining the information of an item or a set of items, conditioned on the respondent's ability (see Equations 6, 7). The fact that the information is conditioned on the ability is central to adaptive tests. Rather than determining a sequence of items in advance, composed to fit all respondents, the sequence can be built continuously as the respondent answers questions. At each stage in this process, the test estimates the ability of the

respondent, and based on this estimate it selects the item that provides the most new information.

In the unidimensional case, implementing adaptive item selection is straightforward. However, a VAA rarely measures a trait that can be captured in one dimension, as the test aims to cover a wide variety of issues. Therefore, an approach that incorporates multiple dimensions is needed. Whereas true multidimensional models exist, they tend to introduce scales that are difficult to interpret, as each item is associated with all scales. To incorporate multiple dimensions and yet preserve the interpretability of a unidimensional scale, the paper outlines a multiscale political space, where each item is associated with only one scale. The scales can then be interpreted in terms of their item set.

A commonly used political landscape is measured on a two-dimensional plane, where one dimension represents a left-right economic scale and the other dimension a social attitude ranging from socially liberal to socially conservative. In this setting, each item is associated with either economic attitudes or social attitudes. The choice of item is now more complex than it is in the unidimensional case, because we are not only comparing the information that the items provide conditioned on an ability. We also have to choose the dimension which the next item is to measure.

In theory, with a multi-dimensional landscape we can still select the item that provides the most information, regardless of which dimension it measures. However, this would not be optimal. If the goal is to decrease the uncertainty in the estimate, we need to consider how much information we already have in each dimension. In a dimension where we already have a low uncertainty, there may be little potential for further improvements even with a highly informative item. Similarly, in a dimension where we have high uncertainty, there is more potential for improvements.

A second reason against applying the maximum information criterion for item selection is that all dimensions may not be equally important. When a user engages with a VAA, the primary purpose is to find out which party or candidate is the closest match. Therefore, we want to focus on dimensions that are important for the ranking of the top contenders. For instance, let a user after several questions be located near two parties while the other alternatives are far off, and let the two closest parties differ primarily in a dimension that we refer to as dimension 2. We are then primarily interested in increasing the precision in dimension 2, as this is most important for the quality of the best-match result.

Paper 1 outlines a method for continuously selecting the best item to present next, taking into account each item's potential to increase precision in a dimension where precision is essential. This allows the user to conclude the test without answering all items, and the matching results will be of higher quality than if the same number of items had been chosen in an arbitrary order. Implementation of the method can make VAAs more available to groups who are prone to survey fatigue.

4.2 Paper 2: Optimized questionnaire item selection for tracking the progression of motor symptoms in Parkinson’s disease

Paper 2 shares a common theme with Paper 1, as both aim at selecting the best subset of items from a questionnaire. In Paper 2, we do so in a longitudinal medical setting where the goal is to efficiently track the progression of Parkinson’s disease over time.

Parkinson’s disease is a neurological disorder that affects movement and motor control. Symptoms include tremor, bradykinesia (slowness of movement), stiffness in the muscles, and postural instability. Non-motor symptoms, such as depression and sleep disturbances, may also occur. The condition develops when nerve cells in a specific part of the brain gradually break down, leading to decreased dopamine production. The disease typically develops gradually, with symptoms worsening over time. There is no cure, but treatments can help manage the symptoms.

The MDS-UPDRS (Movement Disorder Society-Unified Parkinson’s Disease Rating Scale) is an assessment tool designed to evaluate the severity and progression of Parkinson’s disease. It provides a standard for measuring the severity of the symptoms, and by administering the test repeatedly clinicians can track how the symptoms are developing. Monitoring the state of the symptoms facilitates the evaluation of treatments, and allows comparisons across different studies and patient groups.

Each item rates the severity of a symptom or a difficulty caused by a symptom. A symptom is graded on a 5-level scale. For all items that we use in this article, the response options are the same: (0) Normal, (1) Slight, (2) Mild, (3) Moderate and (4) Severe.

The MDS-UPDRS questionnaire has been estimated to take approximately 30 minutes to complete. It has been indicated that use of the UPDRS test is mostly in research contexts (AlMahadin et al., 2020), which raises the question whether a reduced set of test items would be used more often. For this reason, it is of interest to know how well a shortened version of the test preserves the accuracy of the symptom severity estimates.

The MDS-UPDRS test consists of multiple parts, of which some are completed by the patient and some by health care professionals. In Paper 2, we look more closely at items related to motor function. More specifically, we examine 34 questionnaire items that were included in a study by Arrington et al. (2020). This previous study evaluated how the item information obtained from the IRT framework could be used to select items for a shorter version of the test. Arrington et al. (2020) showed that by ranking items based on expected item information, i.e., the Fisher information that we can expect from a randomly chosen individual from the patient population, a reduced data set with 65 percent of the items retained 80 percent of the information.

In our paper, we compared three different methods for selecting an optimal subset of items. An optimal subset, as we defined it, is a subset of items that results in a lower uncertainty about the latent trait than any other subset. The latent trait is here the severity of the motor symptoms. The uncertainty is measured as the expected standard deviation of the latent trait estimate. A key motivation for the paper is that the set of items that maximizes the expected Fisher information for a random patient’s symptom severity is not necessarily the set of items that

minimizes the expected standard deviation. We therefore have reason to explore alternative methods of finding an optimal subset.

The first item selection method that we explored we refer to as coordinate descent local search. The method uses the following algorithm:

1. Select an arbitrary item subset of size K , where K is the number of items that we want in the subset.
2. Go through all K selected items. For each selected item, replace it with each unused item. If replacement with an unused item produces a lower expected standard deviation of the ability estimates, switch the items.
3. After going through all items, repeat the full procedure. Do this until no further substitution produces a lower expected standard deviation.

Below, the algorithm is described in pseudocode.

Algorithm 1: Implementation of coordinate descent local search

Input:

- We have a pool of N items with estimated difficulty thresholds and discrimination parameters.
- We have an estimated population ability distribution.
- K is the number of items in the subset to optimize.
- $subset$ is a vector of K elements, where each element represents an item.

Result:

- A locally optimal subset of K items, where the optimal subset is one that minimizes the expected standard deviation of the ability estimate from a random respondent.

Step 1: Randomly select K items and store them in a vector

$i \leftarrow 0$;

$subset_i \leftarrow \text{select.random}(n.\text{select} = K, \text{from} = 1 : N)$;

Step 2: Substitute items in the vector subset until local optimum is reached.

repeat

$i \leftarrow i + 1$;

$subset_i = subset_{i-1}$;

for $k \in \{1, 2, \dots, K\}$ **do**

$best.sd \leftarrow \text{calculate.expected.sd}(subset_i)$;

$best.item = subset_i[k]$;

for $n \in \{1, 2, \dots, N\}$ **do**

if $n \notin subset_i$ **then**

$subset_i[k] \leftarrow n$;

$this.sd \leftarrow \text{calculate.expected.sd}(subset_i)$;

if $this.sd < best.sd$ **then**

$best.item = n$;

$best.sd = this.sd$;

end

end

end

$subset_i[k] \leftarrow best.item$;

end

until $subset_i[k] = subset_{i-1}[k] \quad \forall k$;

return $subset_i$;

Algorithm 1 is not guaranteed to find the global optimum. However, it will find a local optimum, and by running it multiple times with different initial item subsets we can see whether it ends up with the same result each time. This was the case when we optimized subsets of the 34 motor-function items in the MDS-UPDRS test.

The item subset obtained with the coordinate descent local search algorithm is optimized for the population of test takers, where every test taker answers the same set of items. We also explored an adaptive approach where each patient gets

a tailored item subset, which gives the most information conditioned on the previously estimated symptom severity.

We found that the method of choosing the items with the highest expected Fisher information was an improvement over randomly choosing a subset of items. We gained further benefits from using the coordinate descent local search approach, which directly takes aim at minimizing the uncertainty in the estimate. Whereas the adaptive approach had a slight edge over the coordinate descent approach, it adds further complexity. The choice of approach comes down to making a practical trade-off between precision and simple implementation.

4.3 Paper 3: Estimating Abilities with an Elo-Informed Growth Model

Intelligent Tutoring Systems (ITS) use artificial intelligence to improve education in a digital setting. Their purpose is to replicate the benefits from one-to-one tutoring, but without the need for a human tutor. Unlike a human tutor, a digital tutor can economically scale to benefit a large number of students.

A tutor can, unlike a static textbook, get an understanding of the student's strengths and weaknesses, and can adapt the instructions accordingly. The increasing availability of digital machines, such as computers, tablets and smart phones, provides the necessary infrastructure for a development in the direction of digitally supported learning. Better computing power enables more powerful software based on statistical models. Especially subjects in which learning follows a clear progression, like STEM subjects and foreign languages, can benefit from intelligent tutoring.

Whereas an effective ITS consists of multiple components, each deserving its own treatment, this paper focuses on learning analytics, i.e., methods for tracking the progress of the student. The ability to make a good estimate of the student's current state is necessary to personalize interventions like instructions and practice exercises.

Here, we treat the ability of a student as a one-dimensional latent trait that we want to follow as it changes over time. First, we look at the Elo rating system, originally developed to rate chess players, as a candidate method for ability tracking. Using Elo estimates in education is natural, since Elo chess ratings are in effect ability estimates that update based on the outcomes of one or several chess games.

While the Elo rating system has major advantages as a simple and computationally inexpensive method, it also has limitations. In Formula 22, we see that each update has a step size controlled by the hyperparameter K . A smaller K implies that the gain or loss after a game is smaller, and a larger K implies that the gain or loss is amplified. The choice of step size K is associated with a bias-variance trade-off, especially when we see rapid changes in the true ability, as illustrated in Figure 7.

When we use a small K , the Elo estimate follows the true ability reasonably well as long as the true ability only changes slightly, as in the early timepoints. However, once the change in true ability becomes steeper, the small updates of the estimated ability are insufficient to keep up. We refer to this phenomenon as Elo lag. With a sufficiently large K , the estimate is able to keep up with rapid ability increases, but as shown in Figure 7, this comes at the price of volatile estimates. With an optimized step size, we will have some combination of Elo lag and volatility.

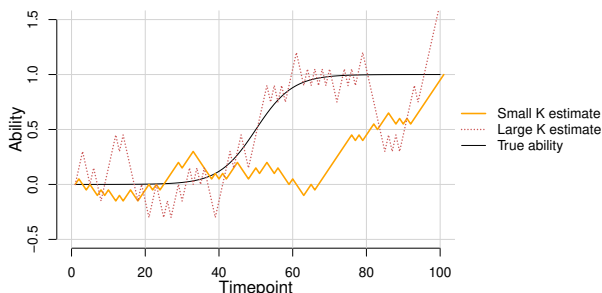


Figure 7: Estimated vs. true ability with different values of K

To remedy the estimation difficulties related to the choice of K , we propose a model that is able to track the ability of a learner without suffering from either Elo lag or high volatility. This is possible by allowing the ability estimation of individual respondents to take advantage of group-level information.

The proposed model, referred to as the Elo-informed model, rests on the assumption that the abilities of a respondent group at any time point follow a distribution that we can estimate. A time point refers to the position in the sequence of items that a respondent answers. For instance, time point 1, also called iteration 1 in the paper, is the time point at which a respondent answers their first question. In other words, the timepoints are discrete and individual, so time point t for two respondents may be far apart in chronological time. Further, the chronological time between time point t and time point $t + 1$ can vary widely from one respondent to the next.

In the paper, we have assumed that the respondent abilities follow a normal distribution with a mean and a variance estimated from the data. A normal distribution is often a good fit for modeling traits that depend on a large number of factors. This is certainly the case with an ability, which depends on numerous inherent traits and environmental factors.

Although the Elo method struggles to track a rapidly changing ability, it works well to rank respondents. Unlike the quality of the actual Elo estimate, the quality of the ranking is not sensitive to the choice of the step size K . The Spearman correlation, which measures how well the ranking of the ability estimates corresponds to the ranking of the true abilities, confirms that the ranking does not suffer much from Elo lag, as this lag affects all respondents.

The algorithm to fit the Elo-Informed Growth Model includes the following steps:

1. Obtain the Elo estimate of respondent abilities over time.
2. At each time point, establish the respondent ability ranking.
3. Based on the ranking and a group-level family of distributions, find the distribution at each time point that maximizes the probability of the observed data.

The model consists of the optimized distribution at each time point, along with the Elo estimates. With this information at hand, we can use the model to estimate the ability of a new user at any time t . We do this as follows:

1. Calculate the standard Elo estimate of the new respondent at time point t .
2. Determine where the new respondent stands in the ranking, compared to the respondents in the training data. Express this in terms of the proportion of the respondents in the training data that have a ranking below the new respondent.
3. From the ranking, along with the group-level distribution at time point t , we can derive the final ability estimate.

To evaluate the model, we used the FIDE ratings of chess players. The FIDE ratings are in effect an ability estimate based on a large number of played games. We used the official rating as our truth, and explored how well the model could reproduce the official rating based on far fewer observations. Using fewer observations, further apart in time, is equivalent to recreating a scenario with more rapid ability changes, as more learning can take place from one time point to the next.

Since the official Elo ratings are themselves estimates, we also conducted a simulation study. In the simulation study, we could validate the performance of the model against a known truth. The simulated scenarios also allowed us to test different values of the hyperparameters against multiple scenarios with varying growth patterns and training data sizes.

In the comparisons, we included three methods: The Elo-Informed Growth Model, standard Elo with an optimized step size, and a GLMM model defined to allow for flexible growth patterns. The flexibility in the GLMM model was achieved by assigning each discrete time point its own group-level mean.

In the model comparison, we calculated the Root Mean Squared Error for each time point. When the true abilities changed more rapidly, the standard Elo method produced, as would be expected, lagging ability estimates. In this scenario, the Elo-Informed Growth Model performed better than standard Elo. The flexible GLMM model performed similarly to the Elo-Informed Growth Model. However, the GLMM produced errors that varied more widely from one time point to the next. A reason for this was that the GLMM mean estimate for each time point was in effect fitted individually, with no information borrowed between timepoints. This makes the estimate highly dependent on a large training dataset. In contrast, the Elo-Informed Growth Model shares data between timepoints, producing smoother growth curves even with a small training dataset.

With slower growth rates, standard Elo with a small step size K was a good option. The Elo-Informed Growth Model and the flexible GLMM model performed equally well. However, the GLMM model still produced errors that varied widely from one time point to the next.

In addition to accuracy, we also evaluated model fitting speed and the need for training observations for each model. With the chess training data, it took the GLMM model about 23 minutes to fit the model. The standard Elo method, on the other hand, requires virtually no time as there is no model to train. Whereas we may run a logistic regression on the training data to find a good starting value for the Elo algorithm, this is done in a fraction of a second. The Elo-Informed Growth Model took 11 seconds to train. It is thus much faster than training a flexible GLMM model, due to a less complex estimation algorithm. Whereas the GLMM model fits all parameters simultaneously, the Elo-Informed Growth Model fits the group-level ability distributions, each associated with one time point, one at a time. This transforms one large task into a series of smaller tasks.

We also examined how long it took to estimate the ability of a new respondent at some time point t . For practical implementation, this time is far more important than the model-fitting time; the model is fitted once and possibly updated once in a while, but estimating individual abilities is a task repeated after each given response. Estimating a new respondent ability did not take more than 0.2 seconds with any of the compared methods.

We found that the Elo-Informed Growth Model is a practical choice when the ability change is rapid, or when we do not know in advance what the ability growth curves may look like. However, if we have reason to believe that the ability growth will be modest, the simpler and computationally less expensive standard Elo method is a natural choice. The GLMM model, as specified in the comparison, would be a reasonable option only with a big training data size. However, GLMM encompasses a wide family of models, and differently defined GLMM models may perform better in terms of accuracy with smaller training data requirements.

4.4 Paper 4: Elo estimated abilities with unknown difficulties

In Paper 3, we compared different methods that can be used to track an ability over time. We assumed that item difficulties were known, which often is not the case. Item difficulties can then be pre-calibrated in a separate process. Pre-calibration can be done with a standard Item Response model using data from a static test setting where a group of respondents answer a battery of items.

However, with limited time and resources, organizing calibration tests may not be feasible. Calibration in a digital environment may require substantial time to build a separate system dedicated to these tests. Administering a pen-and-paper test in a physical venue may be an even more overwhelming task for those who primarily develop digital tools.

An alternative to a separate pre-calibration step is on-the-fly estimation of item difficulties (Klinkenberg et al., 2011). This means that item difficulty estimates are continuously updated with the Elo algorithm. The Elo algorithm is well suited for this, since it is constructed to update the ratings, i.e., the ability estimates, of both chess players after a game. When we go from chess to an educational setting, the respondent is analogous to player one, and the item is analogous to player two. In standard Elo, the gain in estimated ability of the winning player equals the loss in estimated ability of the losing player. The Elo rating system is thus a zero-sum game.

In a closed system, where players neither enter nor exit, the sum of all abilities is constant. The ability updates only redistribute that sum. As a consequence, if we have a group of chess players who practice, so that the average ability in the group increases, this group-level improvement will not be reflected in the ratings. Instead, defending the current rating will grow more difficult over time and demand greater skill.

The zero-sum nature of the Elo rating system can be problematic in an educational context where item difficulties are often viewed as constant. With on-the-fly estimation, we would then expect item difficulty estimates to find the region of the true difficulty, and then vary around this value. This is also the case when the true abilities are held constant over time, as in Figure 8

However, in an educational context we expect respondent abilities to increase

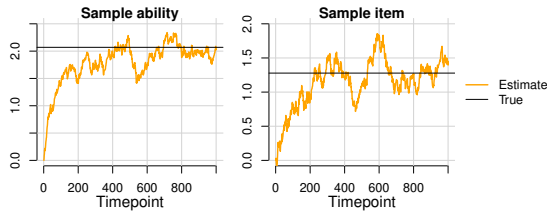


Figure 8: Estimated vs. true values for an ability and an item with on-the-fly item estimation. The true abilities and item difficulties are constant.

over time. In a closed system, this results in steadily decreasing item difficulty estimates. The mean ability may still increase, but the estimates are below the true abilities. The situation is illustrated in Figure 9. This occurs because the gains in true user ability lead to an increasing proportion of correct responses. Each increase in estimated ability will correspond to a decrease in estimated item difficulty. As item difficulty estimates fall, the gains in estimated ability will be held back. Underestimated item difficulties lead to underestimated abilities and vice versa in a cycle. We refer to this systematic underestimation as deflation. The opposite, i.e., inflation in the system, can similarly arise if the true abilities decrease over time. However, this would be a more uncommon situation in a learning context.

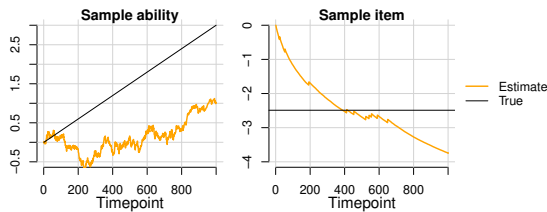


Figure 9: Estimated vs. true values for an ability and an item with on-the-fly item estimation. The true abilities are increasing.

In the Elo rating system, the probability of a correct answer is a function of the difference between the respondent ability and the item difficulty. For that reason, adaptive matching can work well despite deflation in the rating system, as long as abilities and difficulties are equally underestimated. However, allowing overall deflation causes complications. First, it makes comparisons of respondents more difficult. The nature of Elo deflation is that it increases over time. Two respondents, whose activities are separated in time, may objectively have the same ability, and yet the later respondent will receive a lower estimated ability due to the deflation that has taken place in the time elapsed between the two respondents' interactions with the system. A second problem relates to the initial ability of respondents answering their first item. If we, for instance, assign the ability estimate 0 to new respondents, then this will be the basis for matching the respondent to a first item. If the estimated difficulties of the items decrease over time, while the true difficulties are constant, new respondents will be matched with ever more difficult items. To avoid this in a system with deflation, the initial ability of new respondents would

need to be continuously adjusted. For these reasons, we believe that the Elo rating system in an educational setting should measure and offset deflation.

In this paper, we demonstrate that Elo deflation is a real phenomenon, and we propose a method to measure deflation over time. We do this using a step-by-step process:

1. Allow an early batch of respondents to answer items. As we do not yet have calibrated item difficulties, we suggest random matching.
2. Identify items that have been answered several times at iteration 1, and estimate the difficulty of these items. These estimates are made with the Rasch model under the assumption that the group of respondents answering their first item has abilities that follow a standard normal distribution.
3. Calculate the uncertainty of the estimated item difficulties.
4. Calculate the deflation at any time point t as the weighted mean difference between the item Elo estimates at time t and the Rasch difficulty estimates at iteration 1. The weight in the weighted mean is proportional to the precision of the Rasch estimate, so more precise estimates receive greater weight.

For each step, we propose closed expressions as an alternative to numerical optimization, which reduces the computational burden to a minimum while making the method easier to implement.

After calculating the deflation at each time point t , we can offset the Elo estimates. Based on the deflation-adjusted Elo estimates, we can finally obtain a constant difficulty estimate. In the paper, we do this by averaging the second half of the adjusted on-the-fly Elo estimates. The early estimates are thus treated as a burn-in sequence, where the adjusted Elo estimates may not yet have found the region of the true difficulties.

We evaluate the method on a dataset with simulated outcomes and simulated item difficulties. However, to ensure realistic growth patterns, we treated the empirical ability growth of young chess players as the true ability growth. In the results, we could see that after deflation adjustment, Elo estimates of both respondent abilities and item difficulties found the region of the true parameter value and then varied within that region. However, the item difficulties tended to be underestimated when the true difficulty was in the higher end, and overestimated when the true difficulty was in the lower end.

After calculating the constant deflation-adjusted item difficulty estimates, these can be used to estimate the abilities of subsequent users. For comparison purposes, we can also choose to re-estimate the abilities of the early users. The method used to estimate the constant item difficulties does not require the use of any specific model or algorithm in the next step, when we estimate abilities of subsequent users; we can employ any method available for the case in which item difficulties are known or pre-calibrated. An obvious option is to use the standard Elo method, which works well when abilities change at a rate that is not too rapid. Other options are GLMM models, or the Elo-Informed Growth Model that we proposed in Paper 3. With our simulated dataset, using estimated item difficulties, the Elo-Informed Growth Model produced ability estimates almost as good as those produced when the true item difficulties were known.

5 Limitations and Future Research

This dissertation explores methods and proposes extensions that aim to better estimate latent traits. In all included papers, we apply the methods in specific settings and for specific purposes. However, as different contexts can often be described in the same statistical terms, the methods are not limited to these settings.

To analyze the trajectories of motor symptoms in the Parkinson's disease data, we defined a GLMM model to account for population means and individual random effects. We then applied several different methods for selecting optimal item sets. The longitudinal GLMM model and the item selection methods explored in a Parkinson's disease context could also be used to measure many other medical conditions, such as dementia.

The methods we propose all have limitations that we can either accept or use as the starting point for further research. The adaptive IRT algorithm used for VAA data aims to be easily interpretable by assigning each item to only one scale. In contrast, multivariate Item Response Theory models can associate one item with multiple latent traits. As a consequence, the multivariate IRT scales can be more difficult to interpret, but they also make it easier to incorporate items that are weakly associated with many scales rather than strongly associated with one scale. This is a clear advantage when we want to assemble a set of items that cover a wide range of topics relevant to the recommendation. A systematic comparison between our method and adaptive selection that incorporates multivariate IRT would therefore be worthwhile.

Optimal stopping rules for adaptive algorithms have been outside the scope of this dissertation. However, as the aim is to develop a computerized test that allows a lower response burden, it is of great interest to determine when the collection of information is sufficient to conclude the session. In the case of the VAA, we assume that the respondent will choose when to conclude the test and receive a result based on the information provided. In some situations this is acceptable, while in other situations, where the stakes are high, we may only want to provide a result after reaching a predetermined precision threshold.

In the adaptive item selection algorithm, we have not taken into account the advantages of blocking, i.e., the method does not consider that it may be advantageous to group similar items in the sequence. Creating blocks of items allows the respondent to concentrate on one type of item at a time, which may decrease the time it takes to give each answer. How the adaptive selection algorithms proposed here can account for blocks of items is an interesting topic to study further.

The Elo-Informed Growth Model builds on the standard Elo method, which cannot handle ordinal data with an arbitrary number of levels. As incorporating ordinal data would greatly increase the usefulness of the model, extensions that achieve this are worth exploring further. Additionally, the Elo-Informed Growth model assumes that the abilities of respondents at the same time point follow a distribution that can be estimated. This distribution can be the normal distribution, which we used in the model evaluation, or a different distribution that we believe can capture the variation in abilities. A different approach would be to drop the distributional assumptions and instead use the shape of the empirical distribution. This is both simple in theory and computationally inexpensive.

With the Parkinson's data, we track motor symptoms over time while we simultaneously estimate item parameters. These item parameter estimates are used

to optimize subsets of items, based on theoretical properties from Item Response Theory. An alternative method for evaluating optimal subsets would be to fit longitudinal models using different item subsets. The estimates from these models would then be compared to the estimates obtained when using the full item set. In addition, the longitudinal models can be used to forecast motor symptoms at future timepoints. In doing this, estimating the prediction intervals would be of interest as a measure of forecast reliability.

When we limited our study to motor symptoms of Parkinson's disease, we could focus on a single latent construct to track over time. The analysis could be extended to multiple latent states that are dependent. For instance, individuals affected by Parkinson's disease can have both motor symptoms and non-motor symptoms. Estimating these jointly may produce higher precision than estimating them as two independent latent states.

6 Sammanfattning

Latenta variabler modellerar tillstånd som inte kan mätas direkt, såsom mänskliga färdigheter, attityder och hälsotillstånd. Tester som mäter latenta egenskaper i mänskliga sammanhang är ofta utformade som frågeformulär. Utvecklaren av ett sådant test står inför utmaningen att sätta samman en uppsättning frågor som tillsammans mäter ett latent tillstånd med god precision. Vilken kombination av frågor som ger mer precisa skattningar är ett ämne som kompliceras av att det skiljer sig åt mellan olika respondenter. Ett test som är optimerat för att ge god precision för vissa respondenter kan vara mindre precist för andra. Denna avhandling utforskar statistiska metoder för att optimera mätinstrumentet i både simulerade och verkliga miljöer, med olika grupper av respondenter.

Artiklarna som ingår här utvärderar adaptiva metoder som väljer ut frågor baserat på aktuell kunskap om respondenten. Artikel 1 beskriver en adaptiv metod för att välja en fråga i taget i en valkompass. Med ett test som kontinuerligt väljer den mest informativa frågan kan respondenten avsluta sessionen utan att svara på alla frågor, och ändå få ett tillräckligt precist resultat. Den föreslagna metoden bygger på Item Response Theory i kombination med ett multidimensionellt latent konstrukt.

I artikel 2 utvärderade vi en adaptiv modell för att mäta hälsotillstånden hos patienter som utvärderas för symtom på Parkinsons sjukdom. Vi jämförde denna adaptiva modell med optimerade statiska frågeuppsättningar som sammanställts för att ge god genomsnittlig precision över en population. Parkinson-datasetet bestod av upprepade mätningar över flera tidpunkter, vilket krävde ett longitudinellt angreppssätt. I både artikel 1 och 2 var metodernas syfte att möjliggöra mer tids-effektiva versioner av tester och därigenom öka användningen.

Artikel 3 och 4 utvärderar metoder för att följa latenta förmågor som förändras över tid. Till skillnad från Parkinson-scenariot, där ett fullständigt test upprepas vid flera tidpunkter, har vi här endast en observation per tidpunkt. I dessa sammanhang är det vanligt att överge traditionella statistiska modeller och istället förlita sig på beräkningsmässigt enkla algoritmer. Av dessa algoritmer är Elo-ratingsystemet det kanske mest välkända. Detta ratingsystem, utvecklat för att skatta förmågan hos schackspelare, används i dag brett i tävlingssammanhang och även inom utbildning.

Vi har identifierat svagheter förknippade med Elo-metoden och föreslagit metoder för att avhjälpa dessa. I artikel 3 utvecklade vi en hybridmetod som kombinerar Elo-algoritmen med statistisk modellering för att införliva information på grupp-nivå. I artikel 4 visade vi att i ett slutet system där respondenter förbättrar sin förmåga, och där frågors svårighetsgrad uppskattas i realtid, ger Elo-metoden skattningar som deflateras över tid. Vi föreslog en metod för att kvantifiera och kompensera för denna deflation på systemnivå.

References

- AlMahadin, G., Lotfi, A., Zysk, E., Siena, F. L., Carthy, M. M., & Bredon, P. (2020). Parkinson's disease: Current assessment methods and wearable devices for evaluation of movement disorder motor symptoms - a patient and healthcare professional perspective. *BMC Neurology*, 20(1), 419. <https://doi.org/10.1186/s12883-020-01996-7>
- Arrington, L., Ueckert, S., Ahamadi, M., Macha, S., & Karlsson, M. O. (2020). Performance of longitudinal item response theory models in shortened or partial assessments. *Journal of Pharmacokinetics and Pharmacodynamics*, 47(5), 461–471. <https://doi.org/10.1007/s10928-020-09697-x>
- Baker, F. B., & Kim, S.-H. (Eds.). (2004). *Item Response Theory: Parameter Estimation Techniques, Second Edition* (2nd ed.). CRC Press. <https://doi.org/10.1201/9781482276725>
- Bollen, K., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective [Place: US Publisher: American Psychological Association]. *Psychological Bulletin*, 110(2), 305–314. <https://doi.org/10.1037/0033-2909.110.2.305>
- Bolsinova, M., Gergely, B., & Brinkhuis, M. J. S. (2026). Keeping Elo alive: Evaluating and improving measurement properties of learning systems based on Elo ratings. *British Journal of Mathematical and Statistical Psychology*, 79(1), 95–110. <https://doi.org/10.1111/bmsp.12395>
- Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2003). The theoretical status of latent variables [Num Pages: 203-219 Place: Washington, US Publisher: American Psychological Association]. *Psychological Review*, 110(2), 203–219. <https://doi.org/10.1037/0033-295X.110.2.203>
- Cen, H., Koedinger, K., & Junker, B. (2006). Learning Factors Analysis – A General Method for Cognitive Model Evaluation and Improvement. In M. Ikeda, K. D. Ashley, & T.-W. Chan (Eds.), *Intelligent Tutoring Systems* (pp. 164–175). Springer. https://doi.org/10.1007/11774303_17
- Elo, A. E. (1978). *The Rating of Chessplayers, Past and Present* [Google-Books-ID: MeWBAAAAMAAJ]. Batsford.
- Germann, M., Mendez, F., Wheatley, J., & Serdült, U. (2015). Spatial maps in voting advice applications: The case for dynamic scale validation [Makes the point that without scale validation, the items in a set may no measure the same ability.]. *Acta Politica*, 50(2), 214–238. <https://doi.org/10.1057/ap.2014.3>
<https://doi.org/10.1057/ap.2014.3>
- Glickman, M. E. (1999). Parameter Estimation in Large Dynamic Paired Comparison Experiments. *Journal of the Royal Statistical Society Se-*

- ries C: Applied Statistics*, 48(3), 377–394. <https://doi.org/10.1111/1467-9876.00159>
- Klinkenberg, S., Straatemeier, M., & van der Maas, H. L. J. (2011). Computer adaptive practice of Maths ability using a new item response model for on the fly ability and difficulty estimation [Place: Netherlands Publisher: Elsevier Science]. *Computers & Education*, 57(2), 1813–1824. <https://doi.org/10.1016/j.compedu.2011.02.003>
- McCulloch, C. E. (2003). Generalized linear mixed models (GLMMs). In *NSF-CBMS Regional Conference Series in Probability and Statistics* (pp. 28–33). Institute of Mathematical Statistics and American Statistical Association. <https://doi.org/10.1214/cbms/1462106064>
- Mokken, R. J. (1971). A theory and procedure of scale analysis. The Hague, The Netherlands: Mouton. *Mokken A Theory and Procedure of Scale Analysis* 1971, 62(3), 331–347.
- Pavlik, P. I., Cen, H., & Koedinger, K. R. (2009). Performance factors analysis—a new alternative to knowledge tracing. *Artificial Intelligence in Education*, 531–538.
- Reckase, M. (2009). *Multidimensional Item Response Theory*. Springer. <https://doi.org/10.1007/978-0-387-89976-3>
- Sigfrid, K. (2024). IRT for voting advice applications: A multi-dimensional test that is adaptive and interpretable. *Quality & Quantity*, 58(5), 4137–4156. <https://doi.org/10.1007/s11135-024-01845-6>
- Spearman, C. (1904). "General Intelligence," Objectively Determined and Measured [Publisher: University of Illinois Press]. *The American Journal of Psychology*, 15(2), 201–292. <https://doi.org/10.2307/1412107>
- Straatemeier, M. (2014). *Math Garden: A new educational and scientific instrument*
 Doctoral dissertation, University of Amsterdam.