MASARYKOVA
UNIVERZITA

FAKULTA SOCIÁLNÍCH STUDIÍ

Comparison of the Item Parameters Obtained through the Bradley-Terry Model and the Rasch Model

Bachelor thesis

MATEJ RUSIŇÁK

Vedúci práce: Mgr. Hynek Cígler, PhD.

Katedra Psychologie
Program Psychologie

Brno 2024

MUNI
FSS

Anotácia

Táto práca má za cieľ preskúmať možný vzťah medzi parametrami položiek získanými pomocou Bradley-Terryho modelu a Raschovho modelu s cieľom podporiť možnosť merania v psychológii v rámci realistického filozofického rámca. Konkrétne sa zameriava na priame porovnávanie položkových parametrov odhadnutých skrátenou verziou Inventára výšky, ktorá bola administrovaná v dvoch verziách. Prvá verzia predstavovala 4-bodovú Likertovu škálu, ktorá neskôr bola rekódovaná pre analýzu pomocou dichotomického Raschovho modelu, pričom jej vzorka představovala N = 656 odpovedí a tento model vysvetlil 65 % rozptylu. Pre analýzu prostredníctvom Bradley-Terryho modelu boli vytvorené párové porovnania položiek skrátenej verzie Inventára výšky, pričom celkovo obsahovali N = 14,838 pozorovaní a prediktívna sila bola R_{PP} = 0,79. V dôsledku diferenciálneho fungovania položiek (DIF) medzi oboma pohlaviami v analýze Raschovho modelu, dve porovnania odhadnutých parametrov boli vykonané. U mužov, porovnávané parametre položiek výrazne korelovali r_{Adj} = 0,95 a nezávislý výberový test sa ukázal signifikantný t(9) = 7,578, p > 0,001. Všetky položkové parametre, okrem jedného, sa na grafe tiež zoskupili okolo regresnej priamky a spadali do 95% intervalu spoľahlivosti. Podobne aj porovnanie položkových parametrov u žien viedlo k silnej korelácii r_{Adj} = 0,9 a signifikantnému nezávislému výberovému testu t(9) = 6,089, p > 0,001. Zatiaľ čo rovnaký položkový parameter ako v prípade mužov bol vzdialený, väčšina položkových parametrov sa na grafe zoskupila okolo regresnej priamky a spadala do 95% intervalu spoľahlivosti. Tieto zistenia sú však subjektom určitých obmedzení, vrátane malého vzorkového počtu mužov a potenciálnej multidimenzionality Inventára výšky, a sú podrobnejšie rozoberané v diskusnej časti.

Abstract

This thesis aims to explore possible relationship between item parameters obtained using the Bradley-Terry model and the Rasch model in order to support the possibility of measurement in psychology under the realistic philosophical framework. More specifically, it focuses on the direct comparison of item parameters estimated from the shortened version of Height Inventory, administered in two versions. The 4-point Likert scale was administered and afterwards recoded for analysis using dichotomous Rasch model, with collected N = 656 responses, and showing 65 % of variance explained. The pairwise comparisons of the shortened Height Inventory items were created for Bradley-Terry model analysis, with N = 14,838 observations in total and predictive power R_{PP} = 0.79. Due to differential item functioning (DIF) between both sexes in Rasch model analysis, two comparisons between models' estimates were conducted. The comparison of male item parameters yielded a strong correlation r_{Adj} = 0.95 and a significant independent samples test $t(9)$ = 7.578, p > 0.001. All item parameter, except one, were also centered around the plotted regression line and fell within the 95 % confidence interval. Similarly, the comparison of female item parameters also resulted in strong correlation r_{Adj} = 0.9 and a significant independent samples test $t(9)$ = 6.089, p > 0.001. While the same item parameter as in the male sample was found to be distanced, the majority of item parameters were centered around the plotted regression line and fell within the 95% confidence interval. However, these findings are subject to some limitations, including the small sample size of males and the potential multidimensionality of the Height Inventory, and are discussed in detail in the discussion section.

Table of Contents

List of Figures

List of Tables

Glossary

1PL model	–	1 – Parameter logistic model
2PL model	–	2 – Parameter logistic model
3PL model	–	3 – Parameter logistic model
BTM	–	Bradley-Terry model
CMLE	–	Conditional maximum likelihood estimation
CTT	–	Classical Test Theory
HI	–	Height Inventory
ICC	–	Item Characteristic Curve
IRT	–	Item-Response Theory
JMLE	–	Joint maximum likelihood estimation
s-BTM	–	Scale for Bradley-Terry model
sHI	–	Short version of Height Inventory
RM	–	Rasch model

List of Appendices

Appendices

Introduction

Measurement has always been a cornerstone for many different and unique issues in development of the field of psychology. After all, when we try to estimate some psychological trait, it is not as simple as recognizing this trait's existence and then differentiating it from other psychological traits that could possibly affect our estimation. To better recognize this issue, let's imagine a person's psychic as a basket of fruits. For simplification, let's then imagine, that in this basket there are only two kinds of fruit: apples and pears (Martincová, 2024). If we stopped here, it would be easy to express person's trait of appleness in the number of apples they contained and person's trait of peariness in number of pears they contained. But it is not so with actual mental traits. In their cases, pears can sometimes take up apple's skins, as well as they can resemble their shapes and vice versa. How do we express the appleness and peariness of a person then? Psychometrics resolved this issue by simple solution by establishing approach of: "Let's go and make some assumptions." And this exactly was one of the cornerstones of psychological struggle to attain recognition by scientific community as an independent scientific field.

A small revue to the first half of 20th century, lets us encounter a work by Karl Popper called The Logic of Scientific Discovery. Popper states there that empirical sciences consist of empirical theories producing statements that are testable through methodologically appropriate procedures commonly addressed as observations and experiments (Popper, 1935). Statements describing experience cannot be logically considered as statements derived from theories and that "they can occur in science only as psychological statements; and this means, as hypotheses of a kind whose standards of inter-subjective testing (considering the present state of psychology) are certainly not very high." (Popper, 1935). It is worth noting, that Popper here identifies two important points with which psychology struggled before officially being recognized as an empirical scientific discipline. The fact of personal experience being strongly subjective and qualitative disables it from quantification of phenomena and in turn impairs it from inter-subjective comparisons. The methodological nature of empirical science requires the use of measurement based on logical and mathematical laws which cannot be applied to qualitative phenomena. It is then of no wonder that right here arose the debate of whether psychology is capable of measurement and so of being methodologically a sound empirical science altogether (Toomela, 2007).

History of measurement in Psychology

As was already evident, psychology could attain scientific recognition only after it was capable of sound measurement. Edwin G. Boring (1961) divided the evolutionary process of measurement in psychology followingly:

(1) *Psychophysics* founded in 1860 by Fechner's *Elemente der Psychofysik*,
(2) *reaction time measurement,* firstly used in 1862 by Donders in measuring time which it takes for completion of various mental processes,
(3) *quantitative measurement of learning and remembering* by Ebbinghaus in 1885,
(4) and *measurement of individual differences using mental tests* which was pioneered by Galton in his Inquiries into Human Faculty (1883).

It is of no wonder that people grew mostly interested in measurement of individual differences through use of mental states as they provided ability to quantify traits observed in everyday interactions which hitherto were explained only by folk psychology. First pioneer of this was sir Francis Galton, who started within the study of heredity of physical traits where he firstly observed tendency of extreme values towards mediocrity (1886). This subsequently led him to the establishment of normal distribution within heredity and formulation of regression to the mean principle, and thus the regression itself (Galton, 1889). Later, he also noticed some similarities in their respective deviations (Galton, 1883, 1908; Stigler, 1989). It was this that finally led him to the invention of the correlation coefficient through the plotting of a regression line which in turn made it possible to mathematically express relationship between two variables (Bulmer, 2003; Stigler, 1989). His interest in studying talents in people connected with his experience with heredity and advancement in stochastic tools led him to the creation of the first intelligence test which represented the start of psychometrics. This legacy was continued by Karl Pearson who undertook an effort to perfect what he called the Galton coefficient (see Everett, 2010) and based its equation on variance. This made it possible to calculate a correlation for linear relations among variables without need of standardization as it was with the Galton coefficient (see Blyth, 1994; Pearson & Filon, 1898). Just seven years later, Charles Spearman pioneered a correction for attenuation of correlation coefficients which accounted for lowering of coefficients due to the presence of measurement error. As it is later noted in Novick (1966) and a couple years later in Lord & Novick (1968), who finally axiomatized the Classical Test Theory (CTT) as a complex theoretical system for mental test evaluation (Traub, 1997), the discovery of this effect and its correction opened up the way for the concept of reliability and so led to the creation of CTT itself (Trafimow, 2016). It is for this reason, that Spearman is now commonly considered the father of CTT.

Classical Test Theory

Classical Test Theory is laid on two central concepts; true score and measurement error variables (Steyer, 2001), both of which were at least partially elaborated by Spearman. In his writings from 1910 (Spearman, 1910) we can already see some resemblance of observed score being the sum of true score and measurement error, as:

$$X = T + E$$

which became the central mathematical representation of CTT. With the concept of reliability already existing thanks to the advance in correlation theory, Spearman and Brown, independently of each other came to conclusion that reliability increases in relation to lengthening of test, in turn, this helped mathematical representation of reliability (Traub, 1997).

This simple formula provided ability to quantify mental phenomena and so CTT's principles, although not explicitly stated, started to be largely used in test development and score evaluation (Hambleton & Swaminathan, 1985). Around this time, the concept of validity entered psychometrics. In 1921 Buckingham described validity of the test as property of a test measuring what it is supposed to measure (Buckingham, 1921). In other words, we can also call it a problem of a true score of a given test and its relation to what it is supposed to represent. Lissitz & Samuelsen (2007) divide the development of validity into four distinct periods depending on prevalence of philosophical background:

(1) Pre-Cronbach era – rooted in operationalism,
(2) Era of construct validity – rooted in logical positivism,
(3) Era of unified construct validity – rooted in constructivism,
(4) Era of realistic conception of validity – rooted in realism.

By retrospective analysis, we can see that first three eras can be placed within the scope of anti-realist philosophy. And although they didn't consecutively follow as mentioned, but rather overlapped each other and coexisted, this slow turn towards realism is important. From the point of philosophy, realism stands for scientific statements that somehow relate to reality existing independently of our knowledge. This makes it possible to have them confirmed as true or rejected as false at some point in space and time (Dummett, 1982). Anti-realism (and in part instrumentalism) on the other hand is concerned with theoretical description of observable phenomena and stands regardless to those unobservable (Chakravartty, 2017). The difference here is mainly epistemological; realism says that our theories are describing the real world and can be either confirmed or rejected, whereas anti-realism holds that our theories can describe only observable phenomena, meaning phenomena within the scope of our knowledge, irrespective of what the true reality is. Indeed, true reality is of no concern to an anti-realist.

Thurstonian model

American psychologist L.L. Thurstone under influence of operationalism went in a little different direction. His interest was measurement of attitudes and so he developed a method established in the law of comparative judgement which he himself formulated (Thurstone, 1928). He correctly recognized that there are present many dimensions in peoples' opinions towards certain matters. And by isolating one of those dimensions and separation between two stimuli we get an attitude that can be readily measured on a linear scale (Thurstone, 1928, 1954). Construction of such a scale involves isolating statements that relate to a measured attitude and their organization by independent judges by ordering them from those being totally opposite a measured attitude to those being most aligned with the attitude using a frequency distribution (Thurstone, 1928). This frequency distribution is according to Thurstone also symptomatic of the distances between the statements, which he implies, should be operationally close to uniform. If there are then any two statements close to each other, only one of them should be used. Those similar distances then can become units or etalons of measurement. After this is secured, Thurstone remarks that there remains only validation to complete the attitude measure. He mentions three conditions needed for this: a) *The scale must transcendent the group measured*, which means that the order of statements on the scale must be irrelevant of personal opinion of judges sorting it, b) *Objective criterion of ambiguity is observed*, which means that the attitude interval between disagreement and agreement with certain statement is similar for all statements, and c) *Objective criterion for irrelevance*, which covers the fact, that if there is some statement more on one side of attitude measure, the person with the opposite attitude should not agree with the statement, because if he does, it means contamination of the statement by the different attitude (Thurstone, 1928). Based on the distinction made earlier, Thurstone falls into the category of constructivism and operationalism both of which are associated with logical positivism (Bickhard, 2001; Chang, 2021) and he can be claimed to belong into anti-realism realm.

Special case – Bradley-Terry model

For Thurstone's method of comparisons, it was essential to have judges set all items in place respectively (Thurstone, 1928). The Bradley-Terry model (BTM) brings the possibility of an unequal number of comparisons across the items and so in Thurstone's domain presents the possibility of having a larger variety of judges assessing fewer elements. This can in turn better represent the general population. But this is not the primary reason for BTM's invention. The Bradley-Terry Model was firstly discovered in year 1929 by German mathematician and logician Ernst Zermelo in his work *Die Berechnung der Turnier-Ergebnisse als ein Maximum-Problem der Wahrscheinlichkeitsrechnung* as a solution for the then used round-robin tournament format for chess competitions, which required each player to play against each a set number of times. Zermelo's proposed paired comparison model erased the need for the same number of matches played and also solved a problem with relative comparable strength, where the strongest player's estimate would at some point be roughly double the weakest players' indices (Zermelo, 1929). This work was unfortunately largely forgotten, and it was only decades later that it was rediscovered by R.A. Bradley and M.E. Terry in work *Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons* (1952) and subsequently applied to competitions, and was then subsequently applied to various fields, bearing the name the Bradley-Terry model (Glickman, 2013). Later years saw growing interest in pairwise comparisons, and it therefore secured a stable place in applied mathematics (Caron & Doucet, 2012). As of today, BTM is widely used for analyzing paired comparison data as it estimates probabilities of one element being preferred over another (Caron & Doucet, 2012; Matthews & Morris, 1995).

In describing the Bradley-Terry model's central function and structure I will closely follow Cox (1970) and Atkinson (1972). Main function of the model is to denote probability of A_i being preferred to A_j by π_{ij}, while it assumes that elements are ranked only in one dimension and so the π_{ij} depends solely on the difference between ability parameter ρ_i of element A_i and ability parameter ρ_j of element A_j so it equals ρ_i - ρ_j. This outcome is treated as independent Bernoulli random variable with Bernoulli distribution (ρ_{ij}). This makes it advantageous over simple probabilistic model which is heavily dependent on specific elements entering comparison (Fan, 2018). The simplest representation of log-odds from this function is:

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \rho_i - \rho_j$$

Then to solve probability of A_i being preferred to A_j, we get:

$$\pi_{ij} = \frac{e^{\rho_i-\rho_j}}{1-e^{\rho_i-\rho_j}} = \frac{e^{\rho_i}}{e^{\rho_i}+e^{\rho_j}}$$

However, this model is over-parametrized as there are then only t-1 independent values since parameters occur in pairs. We can solve this by prescribing one element entering the model ability of 0 ($\rho_i \equiv 0$), and so make it a useful reference. Every next element entering the model, will then have log-odds of beating $\rho_j = \rho_i - 0$. To estimate the abilities of other elements, the maximum likelihood is usually used which is based on a simple iterative procedure (Hunter, 2004; Zermelo, 1929). In this procedure, element abilities are rescaled after iteration to the point of no significant change in maximum likelihood (ML) (Whelan & Klein, 2022). The result of this process is then an asymptotically standard normal distribution (Atkinson, 1972). With mild assumptions of having independent comparisons and every element presented with a score, while also having some variability in the data, it is indeed a very simple but useful model.

Item Response Theory

Item-Response Theory (IRT) is a complete deviation from CTT modelling era. There are many reasons for the preference of IRT over CTT as I will discuss on next pages but they can be summed up as limitations of philosophical and mathematical background (Borsboom, 2005). The roots of IRT go back all the way to Thurstone and the assumptions of his method that there is a continuous scale underlying responses which allows distinctions of people and items as I will present later in the text (van der Linden, 2010). Based on this idea of some existing scale underlying measured attributes, Paul

Lazarsfeld presented the concept of the latent (unobservable) variable, the level of which is tied to certain responses on an item through some probability (Lazarsfeld, 1950).This connects the latent variable to performance on each individual item and the resulting test score becomes a function of person's ability and item's difficulty that places the person on a continuum of this latent variable (Hambleton & Swaminathan, 1985). If someone answers all of the items on a given test correctly, it doesn't imply that they are at the highest level on the latent variable, but it rather means that item difficulties were not high enough to cover person's ability level. More difficult items can be added so as to distinguish the person's location. In comparison, CTT doesn't distinguish item difficulties and a test score is rather a sum of the correct responses on test items. This means that a person correctly responding to items of what is perceived as the more difficult half of test but due to many reasons, e.g. overthinking, failing items of the less difficult test part will have a resulting score the same (and therefore also level of the trait) as a person who responded vice versa. It is also hard for CTT to cope with a person scoring all items correctly as it creates a ceiling effect for all such individuals (Bjorner, 2019). It is difficult then to add new items to the test to cover those individuals as the test will behave differently and will need a new validation study while also making it impossible to compare individuals across the versions largely due to differences in representative samples (Bond & Fox, 2007). Therefore, CTT operationally states that scores in between the tests are comparable only if condition of parallel tests is satisfied and we are thus comparing the functioning of those tests. IRT is more embedded in realism in a sense that it tries to find a numeral structure that best approximates the underlying trait. Hence the scores that flow from IRT analysis are comparable between any tests as long as they measure the same latent variable (Embretson & Reise, 2000).

IRT is then accurately described as a probabilistic model of latent measurement variables that expresses itself in test items (van der Linden, 2010). It was firstly coherently stated by F. Lord in his dissertation in 1952 (F. Lord, 1952) and by its advancements in 1970s and 1980s strived to replace CTT in test analysis, although it didn't happen completely (van der Linden, 2010). But it is not only the latent scaling that is different in IRT. It allowed an estimate of one's likeliness to respond correctly to more variables (Hambleton et al., 1991). Here I will present 3 similar but different models from the logistic family of IRT models. If we look at the simplest of them which includes only item difficulty and therefore covers person's ability to answer item correctly based solely on item location, we look at 1PL model with structure:

$$P(X=1|\theta,b)=\frac{e^{D(\theta-b_i)}}{1+e^{D(\theta-b_i)}}$$

Where P is an S-shaped curve over the ability scale with values between 0 and 1, θ represents ability and b_i represents difficulty parameter which is located on a scale where probability of correct response is 50%. D (=1.749) in the equation means scaling factor, which makes the logistic function close to the normal ogive function (=cumulative distribution function of normal distribution) (Savalei, 2006). By adding a_i we include item discrimination which affects slope incline and henceforth presents the item's capability of precisely discriminating between people with ability lower and higher than the level needed for correctly responding to item. This results in 2PL model with structure:

$$P(X = 1|\theta, b, a) = \frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}}$$

Within the educational environment, it is also very common to include c_i into the equation which presents parameter of guessing that covers plain guessing of the correct answer when respondent is not certain of any of the answers. This assembles 3PL model with structure:

$$P(X = 1|\theta, b, a, c) = c_i + (1 - c_i)\frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}}$$

In closer evaluation we may notice that modelling structure of all presented logistic models is very similar. In reality, if we take 3PL model and substitute c_j for 0, we get 2PL model and if we set a discrimination parameter $a_i = a$ we get simple 1PL model (Brown et al., 2015).

Rasch model

Rasch model posits a special case of 1PL IRT model with responses quantified in logits, while illustrating connection between the person and an item location on a scale, with the person location being commonly referred to as person's ability and item location being item difficulty (Alfaro-Díaz et al., 2023; Andersen, 1973). This model was developed by Danish mathematician Georg Rasch (1960), around the same time as Lord also ventured into new measurement method centered around a latent variable. Because of the theoretical and mathematical similarities between the theories, RM is considered a special case of 1PL IRT logistic model with one difference from the classical model, which I presented in previous section. RM is calculated with scaling factor of 1 instead of 1.749, which secures its natural logistic ogive slope for Item Characteristic Curve, rather than one resembling cumulative normal ogive (J. M. Linacre, 2005). The equation for primary dichotomous RM, with β_n denoting person ability and other indices identical to 1PL logistic model is then following:

$$P(X_{ni} = 1) = \frac{e^{(\beta_n - b_i)}}{1 + e^{(\beta_n - b_i)}}$$

Estimation of the RM then happens through multiple iterations of item and person statistics through this equation. In the beginning, the person's probability of success is calculated for each person based on the ratio of successfully and unsuccessfully answered items. This is then transformed into an odds ratio and logarithmically transformed. This step repeats until no meaningful change in the score is provided. The same process applies to item locations as well (Bond & Fox, 2007). The average probability of success through logit is set to be 0. This in turn makes it possible for RM to be independent from the sample characteristics and makes it possible for different samples being located in different regions along logit continuum which is for CTT one of the severe shortcomings (Rindskopf, 2001). This is possible only through four central assumptions which RM makes. First, is the assumption of unidimensionality, which just as Thurstone's objection of irrelevance states, that items have single invariant conjoint order and that answering item higher on logit scale correctly also means answering items lower on scale correctly. Second, local independence exists and a response to one item doesn't affect responses to other items. And third, homogeneous discrimination presupposes equal discrimination between people across all items (Gustafsson, 1980). This third assumption is commonly criticized as it posits unreal assumptions for real data. Due to this feature, real-world data never completely adhere to RM. To quantify this adherence between data and RM, fit indices approximated through likelihood estimation are used. As will be stated later in the text, this feature is useful in deciding whether a model is useful in real-world approximation, or it presents a misfit and more extensive theoretical exploration of the latent variable should still be undertaken (Borsboom et al., 2003).

Challenge of measurement

Ferguson committee

Earlier, I introduced some measurement theories and methods that developed in psychology in order to quantify traits. But similarly to validation programs, this sometimes happened independently of each other, and sometimes in a direct response. Although psychology is today regarded as science while hovering somewhere between social and natural sciences, it is preceded by a much larger dispute on its classification. This claim of psychology being a scientific field was not much accepted within the scientific community until the 1930s and 1940s and even afterwards it presents some dispute, although the aim of it has changed. The reason psychology was disregarded as a scientific field was because of the claimed impossibility of sound measurement process in it. This objection was aimed and primarily concerned with psychophysics, which I already mentioned. Measurement of mental processes only followed from advancement in psychophysics which thanks to Weber and Fechner had established some laws of perception, so this objection was not made only against any certain discipline but was meant for any field that strives to measure humans. There was a great unease among the psychologists about it and hence to explore the possibility of measurement in psychology, British Association for the Advancement of Science appointed the Ferguson committee to hold annual sessions starting in year 1932 while having the board consist of psychologists and scientists whose sole aim was to settle this matter.

One of the most influential people on this board was physicist and philosopher Norman Campbell, who already in 1920 coined the fundamental measurement theory (Campbell, 1920) and who supported idea of measurement having resemblance only in physical processes (Campbell, 1932). In his letter concerning the beginning of Ferguson committee sessions he states: "The most likely of these to fail is the associative law that, if a is equivalent to a' and b to b', then a combined with b is equivalent to a' combined with b'" (1932). That means that if we for example take two different pains of comparable strength and two yet other different pains of comparable strength and try to experience them at once, they are most likely to end in different sensations. He acknowledges, that there are people who have made some attempts to measure sensations using other processes different from the ones in fundamental measurement theory, but he concludes that: "They should remember that physicists will not accept any process as measurement, unless it is based upon laws the validity of which is appreciable equally by all observers who are not so abnormal as to fail to appreciate their meaning" (Campbell, 1932). Thus, creating condition that was not to be met in the discourse of that time.

We can sum up objections to measurement in psychology to be engulfed in validity and quantification. While problems of validity concerns mainly question whether we measure what we claim to measure, and were partially addressed by L.L. Thurstone and others of his time, objections on the basis of quantification deal with the fundamental question "Can we measure at all?" or in other words "Are we able to find a unit for our measurement that will be stable and provide properties of the units within physics?" It was for this wording "properties of the units within physics" that disregarded techniques of some contemporary psychologists such as W. Brown, G. Thomson, or L.L. Thurstone who claimed them to be capable to deliver unit-based measurement of phenomena which is carefully distinguished and comprehended. In this era, committee held sessions for seven years but to little avail. The Final Report of the committee stated that it was not convincingly presented that sensation intensities can be fully "represented by a numeral" (Ferguson et al., 1940). But the report concluded that this stance might change due to new discoveries in the future. And in appendix to the Final Report of Ferguson Committee Campbell wittingly states that measurement is to be understood as assignment of numerals to concerned phenomena according to valid rules, i.e. those that can be applied in physics (Ferguson et al., 1940).

New definition of measurement

In year 1946, S.S. Stevens conducted a huge rebuttal of Ferguson Committee findings by establishing a definition of measurement and subsequent properties of the scales used. In his work On the Theory of Scales of Measurement by paraphrasing Campbell (1940), he states:

> *We may say that measurement, in the broadest sense, is defined as the assignment of numerals to objects or events according to rules. The fact that numerals can be assigned under different rules leads to different kinds of scales and different kinds of measurement. The problem then becomes that of making explicit (a) the various rules for the assignment of numerals, (b) the mathematical properties (or group structure) of the resulting scales, and (c) the statistical operations applicable to measurements made with each type of scale* (Stevens, 1946).

He thus breaks with the belief that measurement must be strictly attached to some physical process. By upholding that it is not the possibility of measurement as a whole, but rather the type and properties of measurement, that depend on the nature of measuring unit and its assigning rules, he shifts the discourse to mathematical operations performed after measurement is already conducted. He points out that there are four scales on which measurement can be conducted, ordered from (a) nominal, (b) ordinal, (c) interval to (d) ratio. And their mathematical properties and logical functions are cumulative in a sense that the higher positioned scale possesses all properties of the lower positioned scales (Stevens, 1946). Although he was not the first one, who discovered either of those scales or explored many of their properties, and this distinction was already familiar in the field of physics with examples of ordinal scale in mineral hardness, or ratio scale in temperature measurement in Kelvins, Stevens noted this similarity in representation of the phenomena in physics and psychology, and popularized this thinking in psychometrics (Boring, 1961; Reese, 1943). This helped psychometrics and scientific psychology regain its strength and march towards universal acceptance of CTT in Lord & Novick (1968).

Latest development

Stevens work was upheld and applied in psychometrics but it was not unchallenged (see Michell, 2008). Michell claims, that the scale type selection shifted the attention to the problem of possible transformations that can be performed with given results instead of inquiry of attribute properties. It is important to highlight here, that Stevens was not wrong when he called measurement a procedure of assigning numerals with which even Michell could agree, but the problem rests in incompleteness of such definition (1997). Michell instead, accuses Stevens and psychometricians of malevolency that followed from this denotation. He points out that the much-needed step of connection between the measurement and measured object never followed. Patrick Suppes and Zinnes advanced the theory of this representational measurement by logically connecting measurement and measured attribute (1963). They did this by devising a meaningfulness problem, which implied that a statement coming in a result of measurement is meaningful if and only if its truthfulness doesn't change due to any scale transformation. To sketch this, Suppes proposed two sentences concerning mass measurement: “(i) The mass of the sun is greater than 10, and (ii) The mass of the sun is at least ten times greater than that of the earth. Clearly, (¡) will have a definite truth value only if a particular scale of mass measurement is specified, whereas (ii) has a unique truth value which is invariant under all possible changes of scale” (Suppes, 1969). Truly meaningful, however, is only sentence (ii), which, if indeed true, will be proven so by any scale capable of quantifying mass (Causey, 1969). Michell extends this even further by stating that this is only possible when we have properly quantifiable phenomena, which must be proven before conducting such an estimation (2008). According to him, such investigation is rarely conducted in psychometrics, and hypothesis that psychological attributes are quantitative is commonly without evidence accepted as true (Michell, 2008).

Borsboom, in response to Michell‘s accusation of psychometrics as being pathological, highlights the importance of the implications of the additive conjoint measurement theory proposed by Luce and Tukey (1964). While Michell believes that this theory should be used as a cornerstone of measurement in psychology due to its proper recognition of quantitative properties of the measured attributes, he contends that no such psychometric measurement model adheres to this theory (Michell, 2000). It comes as no surprise for CTT to receive such accusations. As was mentioned earlier, CTT is based on concept of true and observed score and random measurement error, with this error accounting for all the difference between hypothetical true and observed score. CTT, in this regard, becomes infallible and so should be referred to as tautology rather than a proper measurement model (F. M. Lord et al., 1968). Borsboom claims that IRT withstands the objection of an attribute’s quantity assumption. He asserts that there is a distinction between hypothesizing that an attribute is quantitative and therefore deploying quantitative model to analyze it, and directly assuming the quantitative nature of the attribute. According to Borsboom, this hypothesis – that an attribute in question is quantitative – is loosely testable through the fit statistics, which indicate that “IRT models are regularly rejected because they do not adequately fit the data.” However, he also acknowledges that there could be various reasons for this misfit, including cultural differences and item bias, and not solely the problem in the quantitative nature of the latent variable. The impossibility of isolated hypothesis testing is moreover asserted by the Quine-Duhem thesis, which states that testing can be only carried out through a set of hypotheses rather than a single isolated one (Harding, 1976). Simultaneously, under collective ‘hypothesis’ we can include statements regarding qualitative nature of the data, commonly referred to as ‘assumptions,’ alongside statements about the descriptive capabilities of the data (Harding, 1976).

It is interesting to see Michell’s subsequent response to points made by Borsboom. He starts off by stating the cognitive nature of science and leads the way through confirmatory bias, which endangers scientific exploration. He then reiterates from his previous works that psychometrics is blinded by this confirmatory bias of quantitative attributes, and moreover, it deliberately deceives itself through works like Spearman's and Stevens's to dismiss even the slightest doubts. Michell's response to Borsboom's point on model misfit consists primarily of the statement that the quantitative data hypothesis is seldom mentioned and is not even included in IRT constituting works, and therefore, it is doubtful that researchers are even aware of this condition, which when unaccounted for, leads to malpractice. He admits that by directly stating the hypothesis, this psychometric pathology can be erased, although he continues to protest the estimation error and its role in measurement.

Aims of this study

The problem with measurement in psychology can be summarized as one consisting of the uncertain structure of a measured attribute. The history of psychometrics saw the development from psychophysics to mental test analysis and from Classical Test Theory to Item Response Theory. During the first half, it struggled to prove itself as a science by developing sound measuring techniques, and during the latter, it strived to address the shortcomings that accumulated over time and were finally visible only after the change in paradigm (Kuhn, 1962). However, questions about the attribute's nature remained. Some still doubt whether the attribute is indeed real and whether it can be properly quantified. In this work, I will therefore try to contribute slightly to this discourse by connecting Thurstone's operational law of comparative judgment, represented by pairwise comparisons and analyzed through the Bradley-Terry model, with the current state of psychometrics' realist notion, represented by the Rasch model belonging to the IRT family. Based on the reasons mentioned above, I advocate for a realistic approach to measurement in psychology based on the philosophical background of IRT. As previously stated, that according to Suppes' theory proper measurement occurs when different methods provide the same logical values, I will therefore advocate for this realistic approach in psychometrics by attempting to estimate item parameters using two philosophically different but connected probabilistic models. My hypothesis therefore is that: ***the parameter estimates obtained using Bradley-Terry model and Rasch model will not be significantly different when the measurement error is considered.***

Method

Sampling

Respondents were recruited via social media and convenience sampling. Between March 29 and April 9, 2024, INPSY – Institute for Psychological Research – shared a paid advertisement on the social media platforms Facebook and Instagram targeting men and women over 18 years old speaking Czech and Slovak languages. This advertisement, which can be seen in appendix B contained a redirecting link leading to my questionnaire and questionnaire of two other student researchers also covering Likert scale and its respective variations. This collaboration was conducted under grant project called *Response-scale Format Effects on the Psychometric Parameters of Items (SCALING)* supported by Czech Science Foundation (GA23-06924S). Besides, convenience sampling was also used mainly by sharing the questionnaire via student social media groups, but its reach was limited.

The project and this study design were approved by the Research Ethics Committee of Masaryk University (EKV-2022-027).

Participants

After clicking on the provided redirecting link, every participant was taken to one of three Qualtrics questionnaires in the same ratio. The first viewing page of the questionnaire provided a brief explanation of the study purposes, including an informed consent and the possibility of winning a prize of 1,000 CZK if respondent was randomly drawn. The whole description can be found under Attachment B. Description as well as body of questionnaire was written in Czech language and so is also to be found in attachments. After the end of data collection on April 15 I received altogether 907 responses out of which 905 expressed consents with participation in the study. Furthermore, only 675 (75 %) respondents progressed to 97% of questionnaire which was conditional due to identification of respondents' sex and height. Then nine of the participants presented as under the age of 18 and were subsequently imputed. In nine cases respondents filled out the survey twice and only their first answer was used. One case contained blank questionnaire and was therefore also filtered out. In one case, age was written as 0.36 and was changed to 36.

Final sample presented *N* = 656 respondents. Sample contained predominantly female participants with *N* = 550 (84 %) and lower proportion of male participants with *N* = 91 (14 %). Questionnaire was filled out by *N* = 15 (2 %) participants who selected option "other or I don't want to answer" and due to the nature of used inventory were not included in any of the analysis. Twelve female participants didn't fill the column for age. Overall mean age of the sample was 32.36 years (*N* = 644). Female participants' mean age was 32.98 (*SD* = 14.12, *N* = 538) and male participants averaged 30.51 (*SD* = 14.22, *N* = 91). Distributions for age across both sexes were significantly positively skewed but didn't differ much from one another. For this reason, I decided to perform asymptotic Mann-Whitney U test which didn't show significant between-group difference with *W* = 27504, *p* = 0.059. The item for height was not filled by one female and one "other" respondent. Mean height was 169.64cm (*SD* = 8.41, *N* = 654). Female participants reported mean height of 167.78cm (*SD* = 6.99*, N* = 549) and for male respondents it was 180.82cm (*SD* = 7.61*, N* = 91). Distribution for sample in general as well as respective groups was close to normal. Therefore I conducted Welch's t-test for independent samples with male respondents showing significantly greater height with *t*(116.59) = 15.31, *p* < 0.001, *Cohen d* = 1.84. Respondents were also asked to state their highest attained education. 47 of them (7 %) provided elementary education, 31 (5 %) secondary education without Matura, 323 (49 %) accomplished secondary education with Matura, and 255 (39 %) stated they possess some type of college education.

Measures

For this study, I used the shortened version of Height inventory (HI) formerly created by Rečka (2018) and shortened by Tancoš (2019). Original version of this inventory utilized 4-point Likert-type scale. Short version of this inventory consists of 11 statements, six of which are reversely scored, and is based on the representation of one's own height in relation to everyday experience. In the former study, Rečka determined that after transformation of reversely scored questions, the scale creates a unidimensional structure with the underlying construct being hypothesized to be psychological height. As there is difference in physical height between men and women, the scale also exhibits sex as a substantive factor and is therefore suggested to be analyzed independently (Rečka, 2018). Short version of scale according to Tancoš (2019) retains its psychometric properties and was therefore analyzed to primary recommendations.

My questionnaire consisted of three variations of this scale in total, but one concerned with pairwise comparisons in relation to one's experience is outside of scope of this paper.

Agreement

First variation presented the short version of Height Inventory (from now on sHI) with respondents expressing their stance in relation to statements by 4-point Likert-like scale ranging from 1 to 4 (1 = strongly disagree, 2 = disagree, 3 = agree, 4 = strongly agree). Full version of this scale can be found under Attachment A and it is also possible to experiment with the tool first-hand on website (Cígler, 2019). I decided to use Rasch model for item analyses due to objectives aforementioned. For the purpose of using the former dichotomous RM, I recoded answers involving disagreement (levels 1 & 2) as value 0, and answers involving agreement (levels 3 & 4) as value 1 for all positively scored items. Opposite scoring was applied for all negatively scored items (i.e., agreement was coded 0 while disagreement 1). For the item parameter estimation, midpoint where 0 changes to 1 then became important.

Items' rating

Second variation was needed to feed information to Bradley-Terry model and so I selected pairwise comparisons. In these comparisons, pairs were created from sHI items making altogether 55 unique pairs. The position of statements within individual pairs was randomized and this information was later fed to BTM. To be able to estimate the position of items on the logit scale provided by BTM I specifically requested respondents to always choose a statement that is more appropriate for the higher person from the pair displayed. This request was followed by assessment of comprehension through providing participants a pair of statements from HI that are not in sHI and which exhibited a large difference in item difficulty in Rečka's study (2018). After the correct answer was selected, the respondent proceeded then to scale itself. If the wrong answer was selected (9.9%), further explanation was provided and the same pair for comprehension check followed, if the correct statement was selected, respondent proceeded to the scale. If the wrong answer was repeatedly selected (11 %), the respondent was forwarded to next section of questionnaire. Participants answering the scale for Bradley-Terry model (from now on s-BTM), were presented with 25 pairs randomly selected from the pool of 55 unique comparisons. This tactic was selected due to duration of responses to such comparisons and participant fatigue. In case of respondent's indecisiveness between the two presented statements, button "I do not know" was displayed after 10 seconds. This button was delayed, to encourage respondents in selection of one of the presented two statements.

Sample size

Wright states that 500 responses are almost always enough for reliable estimation of item parameters via Rasch model but even sample as little as 100 respondents may be used with awareness of attenuated power (1977). Since my questionnaire required this number of observations for both sexes, while participation in online questionnaires is usually favored by women at around 70% (Smith, 2008), I estimated that I needed around 830 responses overall to collect sufficient data for RM estimation. However, due to Meta algorithm used on their social media platforms, women were targeted more with paid advertisement, as it was cheaper to collect one response from female participant than from a male. This furthered division between sex representation in the sample and for insufficient male representation, I decided to work mainly with data from female participants in RM.

For estimation of sample size for BTM, I conducted power analyses, which consisted of simulation of population of item parameters which were then compared with respective item parameters from Rečka (2018). Analyses showed that correlation of item parameters was at sample size *N* = 50 close to asymptotic, with *r* > 0.995. Nevertheless, this alignment exhibited itself strongly from sample size *N* = 30 and so I decided to consider this as minimum number of observations per paired comparison. Due to presenting only 25 pairs out of 55 (46 %), the minimum sample size had to be multiplied by a ratio of pairs available and pairs provided. This inflated the minimum sample size to 66 respondents under assumption of an even display of comparison pairs. Following figures show comparison of parameters and estimates after every 25 iterations of the power analyses (Figure I) and Recovery plot expressing correlation between the parameters and estimates with respect to sample size (Figure II).

Figure I: Comparison of parameters and estimates

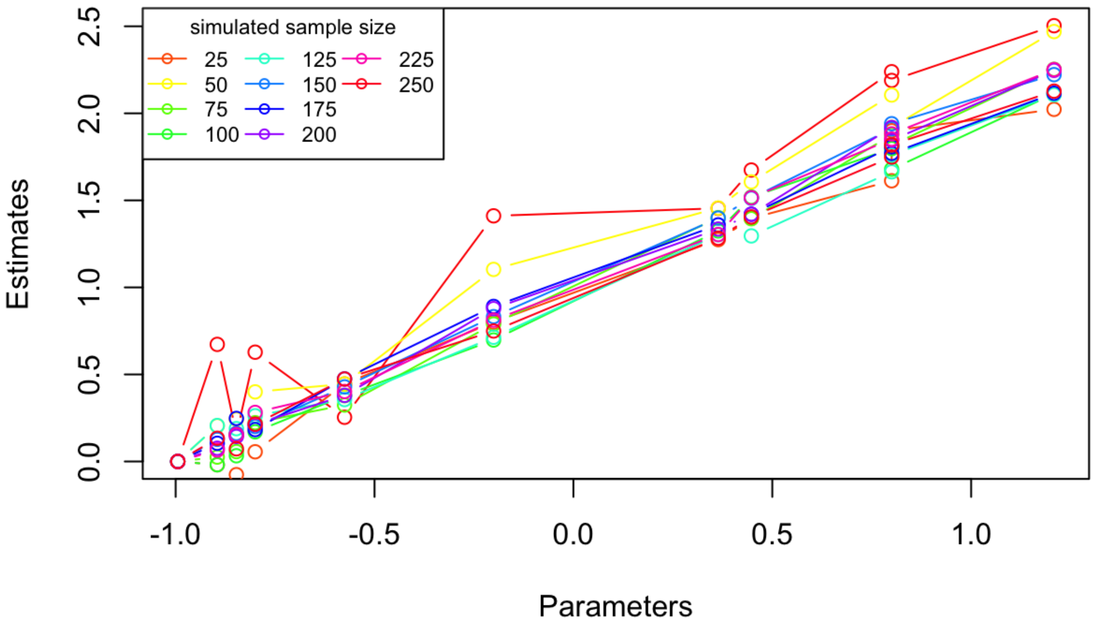
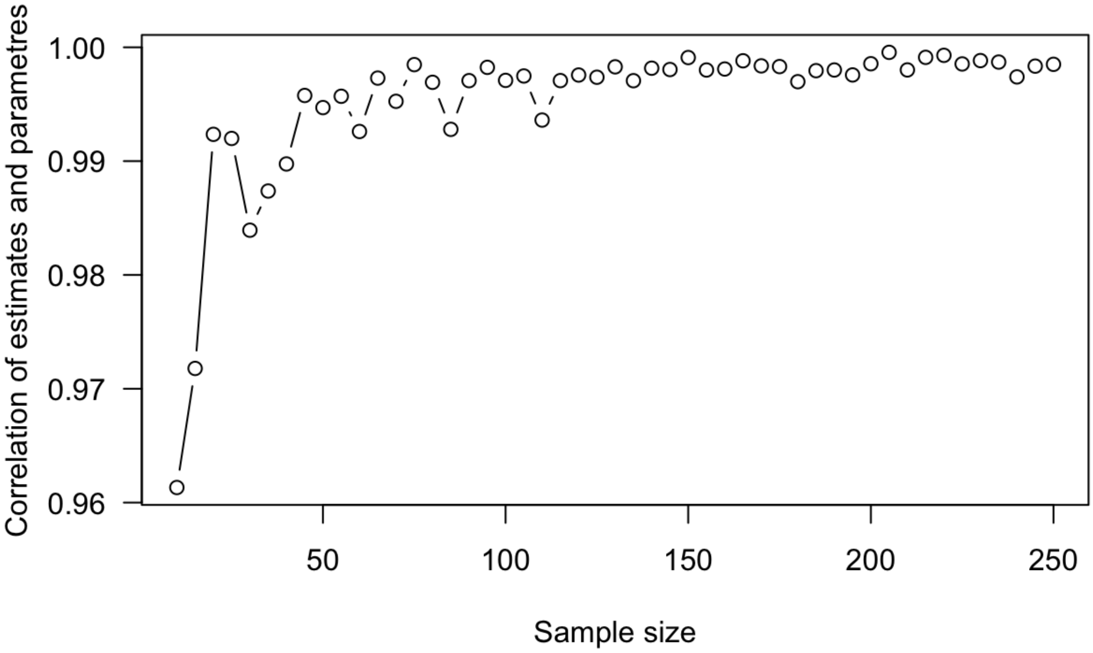


Figure II: Recovery plot from power analysis



Statistical analysis

Dataset as well as code R code used for the thesis is available at Open Science Foundation (OSF) repository at osf.io/umf3h. For the whole analysis, statistical language R version 4.3.3 was used (Posit team, 2024). The basic manipulation of the data and consequent analyses were handled using psych R Package (Revelle, 2024), dplyr R Package (Wickham et al., 2023), and tidyverse R Package (Wickham et al., 2019). For the Rasch model estimation TAM R Package (Robitzsch et al., 2024) and eRm R Package (Mair & Hatzinger, 2007).

Results

Bradley-Terry model

Data was extracted from the paired comparisons responses to Qualtrics questionnaire and prepared for Bradley-Terry model. Data frame consisted of winner and loser column, response column denoting position order of the lower numbered item from sHI. Value 1 meant that item (statement) with numerically lower ID was displayed on the left from the two, and value 0 represented position on the right. This was entered into equation of BTM as response variable to account for preference of items being displayed on left/right. Besides this, sex was also included in the data frame. Items were denoted by their respective IDs and altogether accounted for *N* = 14,838 paired comparisons. Two Bradley-Terry models were estimated with one including moderation through sex (Table 1).

Table 1: Bradley-Terry models

	Combined model		Model with moderation			
			Women		Men	
	Parameter estimate	SE	Parameter estimate	SE	Parameter estimate	SE
(Intercept)			0.026	0.026	0.026	0.026
V1	-0.643*	0.075	-0.671*	0.081	0.197	0.217
V2	-0.437*	0.075	-0.389*	0.081	-0.361	0.214
V3	-0.514*	0.075	-0.493*	0.080	-0.193	0.224
V4	-0.748*	0.075	-0.713*	0.081	-0.272	0.225
V6	-5.615*	0.124	-5.691*	0.136	0.399	0.332
V7	-6.236*	0.130	-6.314*	0.143	0.415	0.347
V8	-4.585*	0.117	-4.646*	0.128	0.336	0.316
V9	-3.699*	0.110	-3.708*	0.121	-0.009	0.303
V10	-4.976*	0.119	-4.970*	0.130	-0.098	0.326
V11	-4.753*	0.118	-4.785*	0.129	0.137	0.321

*p < 0.001

For both models, item V5 was used as reference due to its greatest win-to-lose ratio. That means that in combined model, its estimate β = 0 (*SE* = 0) and its error is spread across the rest of estimates variables. In model with moderation, its value is represented by intercept. Men parameters in moderation model are a resulting product of women estimates and men indices, while intercept staying the same. All estimates for men were non-significant. To evaluate change in the model-fit through residual variance, likelihood ratio test (LRT) was performed with unsignificant result ($\chi^2(11)$ =

4.77, p = 0.058). With correlation between men and women estimates from moderation model being r = 0.99, only combined model without the interaction was investigated further. For the combined model, I calculated R_{PP} statistic denoting predictive power of the model. This statistic is actually a correlation of observed and predicted estimates (Baguley, 2012), which was firstly denoted by Zheng and Agresti (2000). For combined model is this statistic R_{PP} = 0.787, suggesting good prediction. Figure III presents logit-scale estimates for each item, together with 95% CI. We can see a separation of parameters between item V4 and V9 and altogether closer alignment of higher scoring items compared to lower scoring items. This separation can be seen also in Figure IV presenting win-to-loser ratio of all items respectively, and is most likely created due to reverse wording of the items as change in item difficulty for reversed items is also reported in Rečka (2018).

Figure III: Logit scale with item v5 = 0

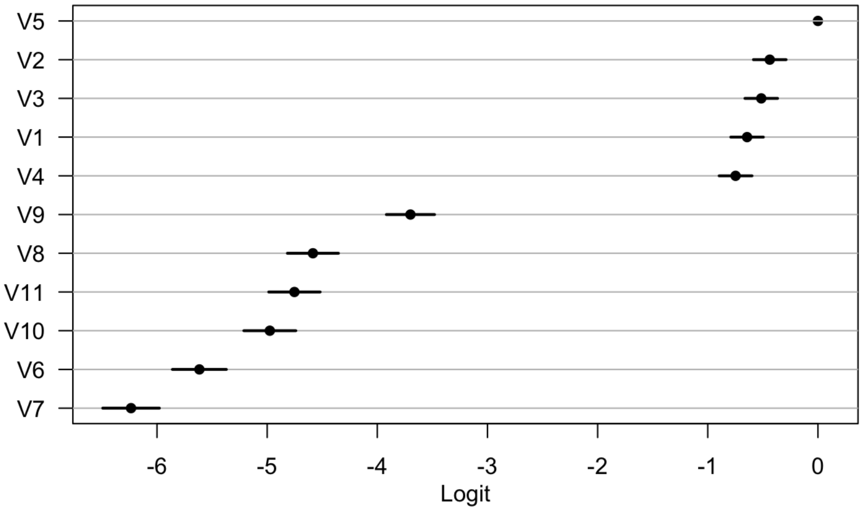
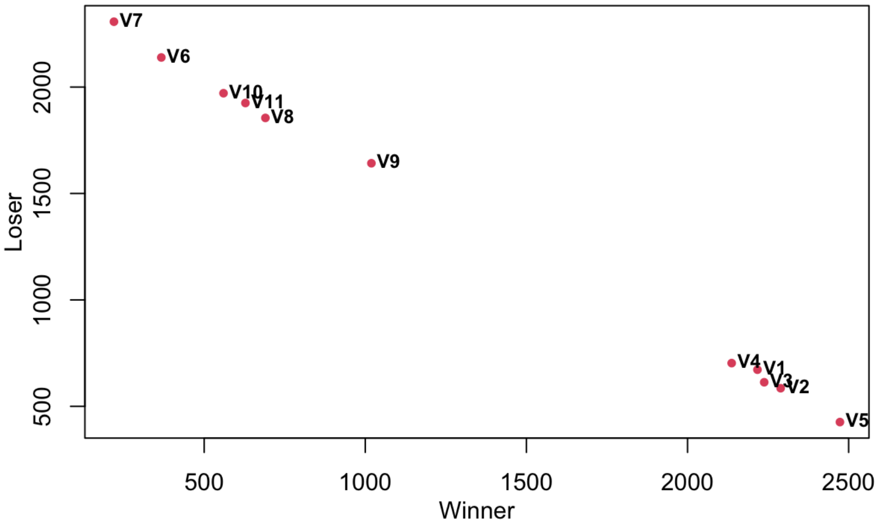


Figure IV: Winner-loser comparison: combined



Bradley-Terry model assumptions

Following power analysis I correlated Bradley-Terry item estimates with CTT popularities that were used for sample size estimation with *r* = - 0.98 suggesting good overall estimation. Sufficient number of comparisons for the BTM in order to prevent misfit of the model was secured by randomized selection of 25 pairs out of 55 possible combinations and those were presented to each respondent. On average, every pair was evaluated 276 (*SD* = 30.82) times with *min* = 208. Enough observations were provided for both sexes respectively. Figure II suggests a good overall representation as the linear spread of wins and losses makes it possible to estimate respective items more precisely and becomes steady ground for transitivity. As in the case of less popular item regularly beating more popular, it would be shown by change in the spread of the items (Wu, Junker, et al., 2022). To check for possible unequal variances across the estimations I calculated randomized quantile residual (RQR) suggested by Dunn and Smyth (2018) for general linear models with binomial distributions. Figure V then shows boxplot of these residuals, and Figure VI presents their Q-Q plot. This supports independence of comparisons and so appropriates model-fit (Wu, Niezink, et al., 2022).

Figure V: Boxplot of RQR

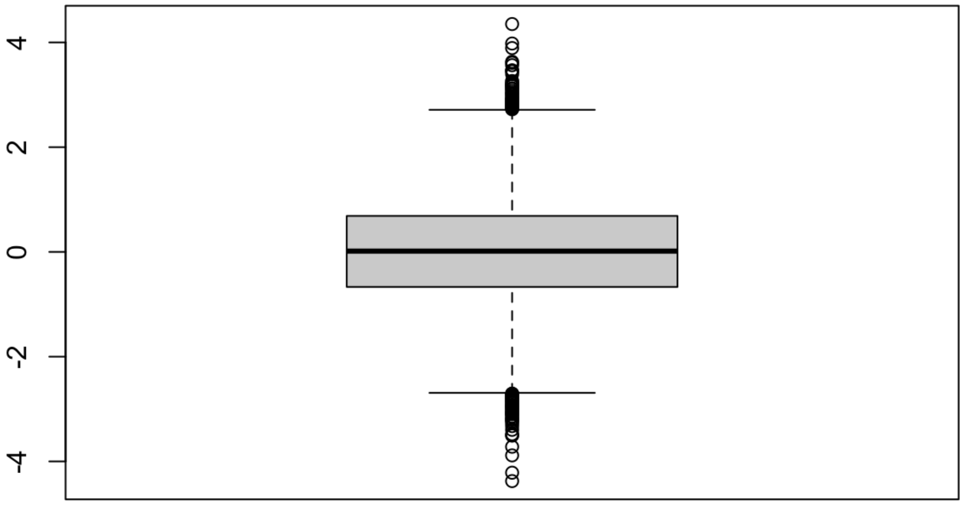
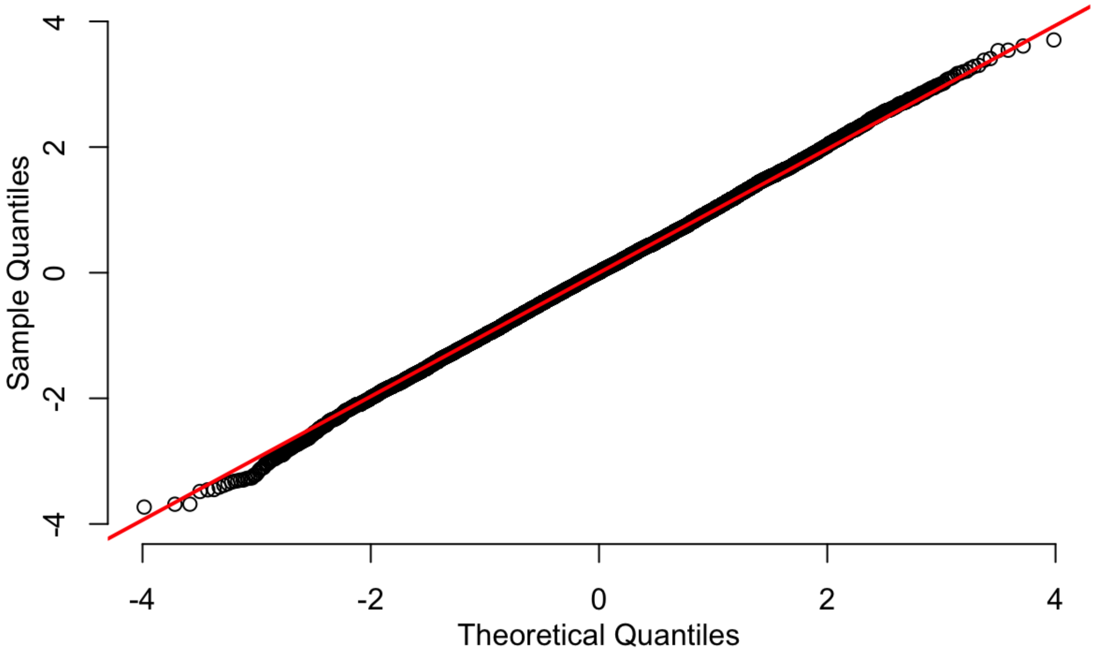


Figure VI: Q-Q plot of RQR



Rasch model

Data from the Likert scale section of questionnaire was analyzed using dichotomous logistic Rasch model (Rasch, 1960). I conducted preliminary analysis using maximum likelihood estimation, which suggested different item functioning (DIF) for men and women, which confirmed findings of the validation studies (Hubatková, 2020; Rečka, 2018; Tancoš, 2019) which observed varying CTT popularities across sexes. Therefore, I decided to perform separate models for each sex. However, while the women's group (N = 550) provides sufficient data for Rasch model (RM) estimation, the same cannot be said for the men's group, with only N = 91 responses collected. This poses a significant challenge for the estimation of the Rasch model and the evaluation of goodness of fit. Therefore, I will provide only a brief summary of the results on their analysis. Women exhibited only 4 missing values in total, three of which were accounted for by just one participant. This participant stated in the comment section that they were unable to refer to some items as they are disabled. However, due to the nature of Rasch model and its ability to pinpoint even participants with the missing data, every participant was entered into the model. In total 45 participants ended up agreeing with each item (after recoding and transformation) and 23 participants disagreeing with each item. Although this doesn't provide model with meaningful variation, estimation package in R studio was already trained for these values.

For Rasch model estimation I used joint maximum likelihood estimation (JMLE) advocated for by Wright and Panchapakesan (1969) which produces the same estimator as method for the for Bradley-Terry model. As in BTM, JMLE completes cycle of estimation of parameters for items and respondents and then enters them back into the model of estimation until no meaningful change is provided. And commonly as in BTM, theoretically is JMLE capable of providing two equal estimates for items as well as for respondents if they have similar response pattern.

Table 2: Model Summary Table: women

Statistic	Items	Persons
Logit Scale Location Mean	0.557	0.000
Logit Scale Location SD	2.378	2.894
Standard Error Mean	0.146	1.024
Standard Error SD	0.017	0.169
Outfit MSE Mean	1.403	1.402
Outfit MSE SD	0.591	4.425
Infit MSE Mean	0.821	0.785
Infit MSE SD	0.126	0.563
Std. Outfit Mean	0.769	0.909
Std. Infit Mean	-2.385	-0.295
Reliability	0.996	0.845

Table 2 displays item calibration for women, using average logit-scale calibrations, standard errors, and model-data fit statistics. Examination of the results indicates that, on average, respondents located lower on the logit scale (*M* = 0.00, *SD* = 2.89), compared to items (*M* = 0.56, *SD* = 2.38). This suggests that the items were a little bit more difficult for this sample to agree with on average. Elevated Standard Error Mean for respondents (*M* = 1.02) in comparison to items (*M* = 0.15) also suggests some issues related to targeting some the respondents. Examination of Table 3 and subsequent Figure VII with joint item characteristic curves shows this issue closer. As average respondent's latent trait logit is 0, there is considerably bigger gap in location logits between item v1 and v11, which could lead to loss of some information although not crucial (Salzberger, 2003). Examination of model-data fit statistics however suggests a challenge in the overall fit to the model. Both, item *Outfit MSE* = 1.40 and person *Outfit MSE* = 1.40 indicate haphazard item functioning coming from unexpected observations of respondents on items that are outside of their ability, either very easy or very hard (Linacre, n.d.). *Infit MSE* statistics on the other hand show a lower variability than expected (item *Infit MSE* = 0.821, person *Infit MSE* = 0.785), albeit not greatly, which means that there were not so many unexpected responses concerning items whose location matched with respondents' ability. In our case, Linacre also suggested, that Infit statistics may be more appropriate to look at when evaluating overall model fit as Outfit statistic is greatly influenced by outliers (Linacre, n.d.; and compare B. D. Wright & Masters, 1990).

Figure VII: ICC for women's data

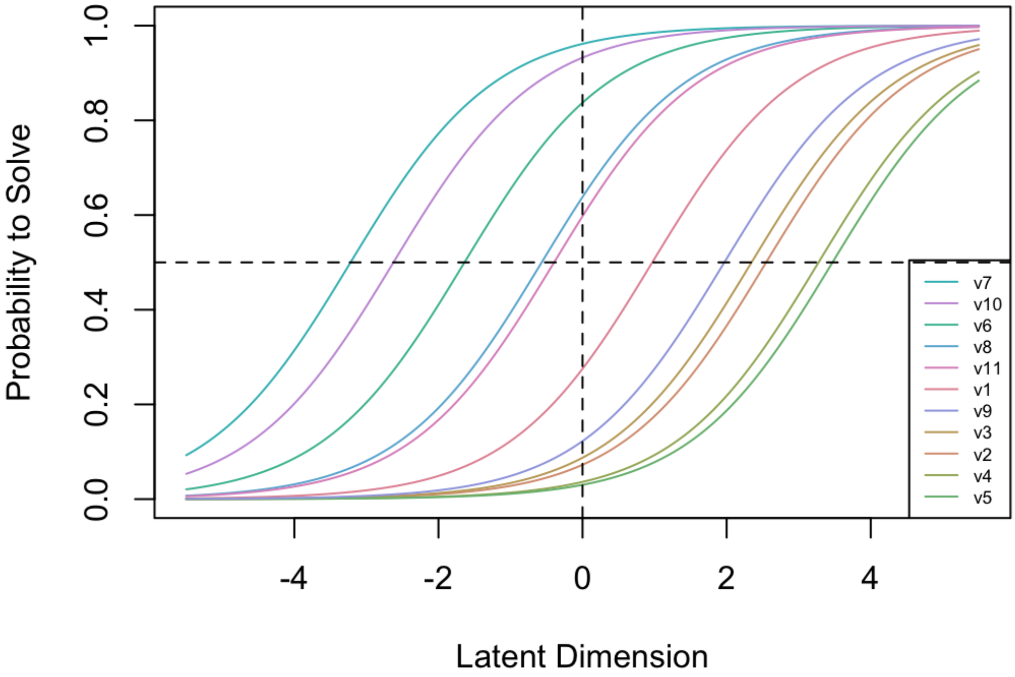


Table 3: Item Calibration Table for women

Task ID	Proportion Correct	Item Location	Item SE	Outfit MSE	Std. Outfit	Infit MSE	Std. Infit
v5	0.138	3.47	0.173	1.295	0.61	0.72	-3.114***
v4	0.151	3.276	0.168	2.041	1.15	0.888	-1.18
v2	0.209	2.548	0.151	0.486	-0.719	0.698	-3.988***
v3	0.227	2.346	0.147	1.525	0.942	0.755	-3.273***
v9	0.264	1.968	0.141	1.828	1.485	1.059	0.78
v1	0.375	0.968	0.129	1.133	0.496	0.79	-3.316***
v11	0.545	-0.395	0.125	2.092	2.859**	0.959	-0.613
v8	0.567	-0.567	0.126	1.371	1.151	0.912	-1.35
v6	0.696	-1.638	0.133	0.428	-1.305	0.73	-4.278***
v10	0.796	-2.624	0.149	2.088	1.296	0.87	-1.628
v7	0.845	-3.219	0.163	1.147	0.493	0.651	-4.276***

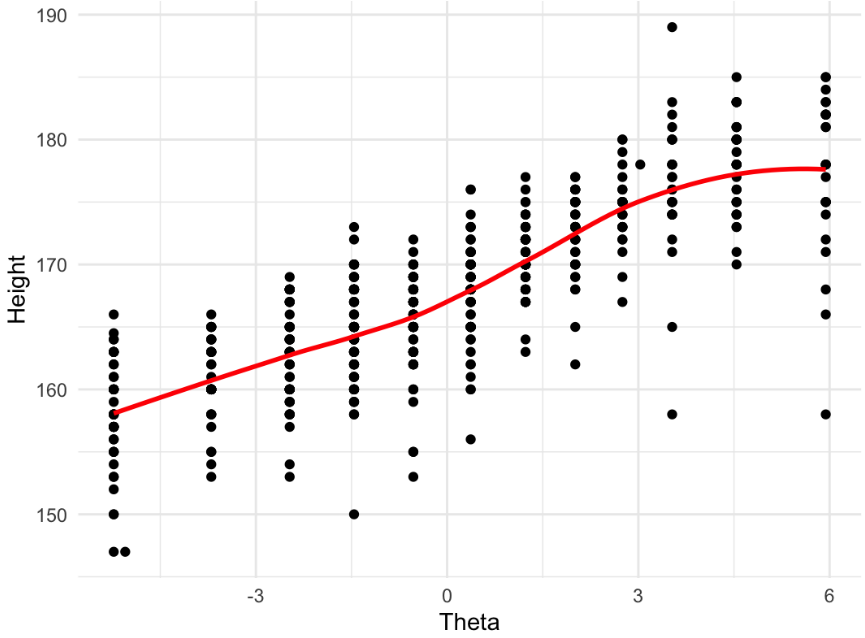
* p > 0.05, ** p > 0.01, *** p > 0.001

For better understanding of the problem, we can look at Table 3 displaying separate item statistics. Items are ordered by their location on logit scale. The hardest item v5 is separated from the easiest item v7 by more than 6.5 logits which means that there is difference of more than 6.5 levels in latent trait between just 11 items. This could possibly affect participants as it makes hard to finely estimate their ability. By Outfit MSE and Std. Outfit we see deviation from the expected observations by the model. Linacre suggests that parameters between 0.5 and 1.5 are productive for the measurement (2002b). As such, we have 4 items complying with this condition. Values from 1.5 to 2.0 he refers to as unproductive but not degrading. And with “degrading” being values over 2.0. This issue however doesn’t affect additivity of measurement, but rather presents a model that distorts reality. Items with larger Outfit MSE values are less discriminative between respondents and therefore introduce greater ambivalence to model (Linacre, n.d.). However, rather than appropriate estimation of person parameters, I am concerned with item parameters which are to be compared with BTM estimates. Therefore, although the model distorts reality, I will use its estimates in comparison with Bradley-Terry model, while being aware of its shortcomings and possible attenuation of correlation between them.

Rasch model assumptions

Due to the nature of construct measured, we can compare estimated latent trait indices with real-world criterion, personal height. Theta to height correlation yielded r = 0.82, which supports validity of the test and complies with Rečka’s hypothesis, that although HI is created to measure physical height, it may be measuring psychological height instead, which accounts for some variation between real height and test measure (Rečka, 2018) and this relationship is shown in Figure VIII.

Figure VIII: Theta and height comparison for women



Raw variance measure shows that this model accounts for 64.78% of variation within the data. To check for local independence, I used Yen's Q3 statistics. It is generally recommended to beware of the coefficients over 0.2 or under -0.2. After adjustment, Q3 statistic presented $M = 0.00$ ($SD = 0.065$) with range from -0.14 to 0.12 supporting assumption of local independence (Quittre & Monseur, 2010).

Rasch model for men

RM estimation for men showed more conservative values in terms of Outfit and Infit statistics, generally suggesting better fit than the model for women. Item reliability showed $r = 0.97$ and person reliability $r = 0.78$, which suggests that the hierarchy of items is easier established than the exact order within within-group environment. Therefore, although people cannot totally agree on the exact order of items, they clearly create similar hierarchies of them. The whole measure was however shifted with Logit location mean for items $\bar{X} = -1.23$ and for persons $\bar{X} = 0.26$, although it is unusual, as JMLE estimation usually provides measure with person location centered around values 0 on logit scale. After closer inspection of joint ICC, however, big gap can be seen between items v1 (β = - 0.11) and item v11 (β = - 2.77), which can result in loss of information between from near items (Linacre). Yen's Q3 statistics showed values between – 0.29 and 0.47 suggesting constraining local dependency (Quittre & Monseur, 2010).

Comparison of item parameters

After the primary analysis of the data through Bradley-Terry model and Rasch model, their respective item parameters were plotted against each other. To retain statistical power of the model, BTM estimates of combined data were used, as they showed between sex invariancy. The plotting included regression line, 95% confidence interval, and 95% prediction interval to facilitate data's ability of prediction.

Figure VII displays comparison of BTM and RM for women. Observed correlation between parameters was $r = 0.897$. For the fact of attenuation of correlation due to reliability estimates for the respective models I used Spearman attenuation formula and corrected correlation was $r = 0.907$. Test for independent samples came also in support of correlation $t(9) = 6.089$, $p > 0.001$, 95% CI of correlation (0.643, 0.973). Subsequently, to evaluate Rasch model I evaluated predictive performance after cross-validation with Bradley-Terry model, which was significant, $\chi^2(9) = 397.46$, $p > 0.001$, and therefore suggested poor performance of the predictive power.

Figure IX: Bradley-Terry and Rash model comparison for women

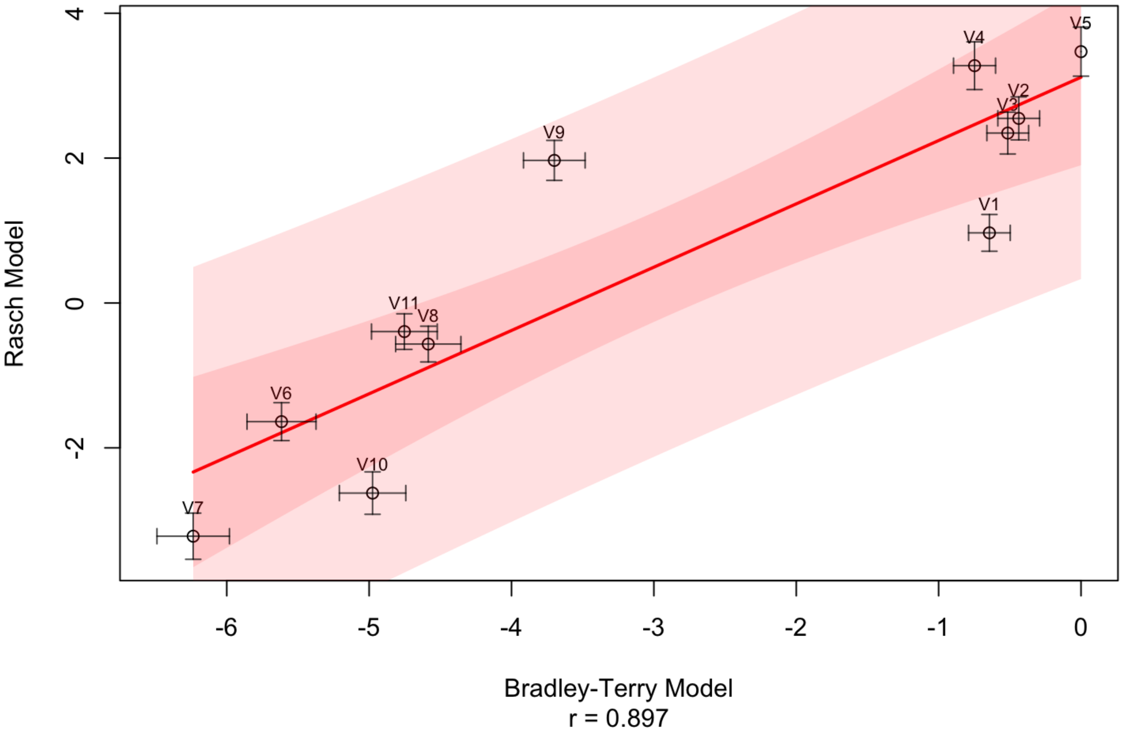


Figure X: Bradley-Terry and Rasch model comparison for men

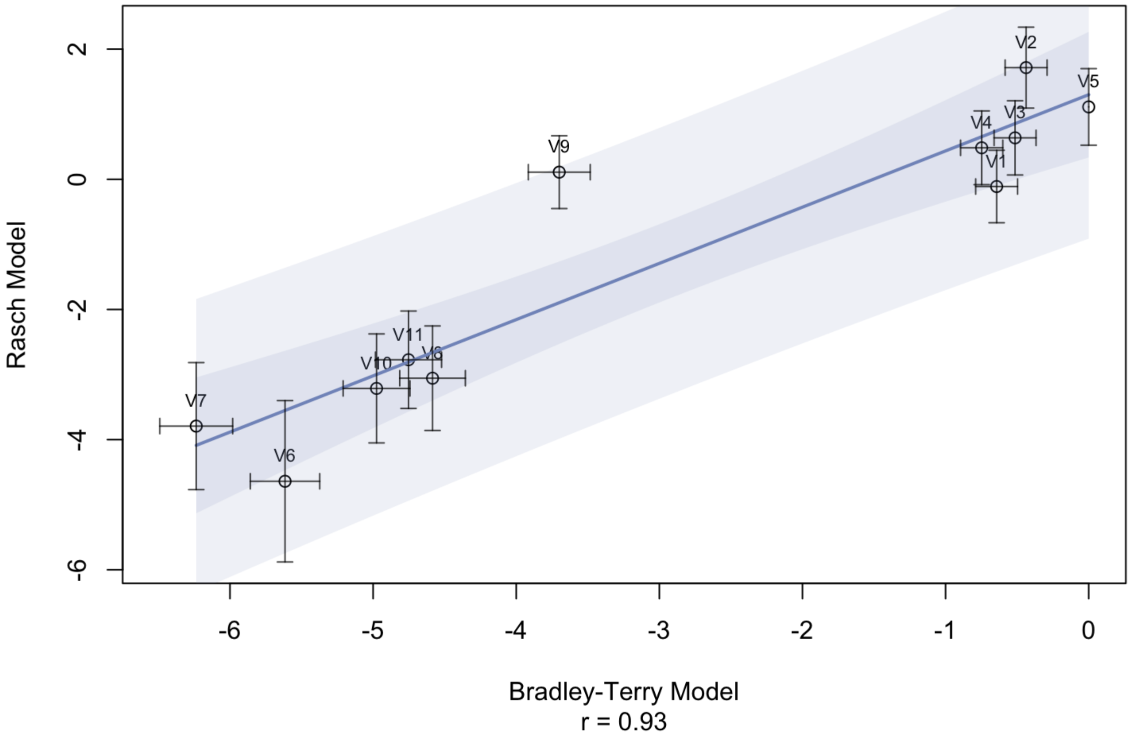


Figure VIII displays comparison of BTM and RM for men. Observed correlation between parameters was *r* = 0.929. For the fact of attenuation of correlation due to reliability estimates for the respective models I used Spearman attenuation formula and corrected correlation was *r* = 0.952. Test for independent samples came also in support of correlation *t*(9) = 7.578, *p* > 0.001, 95% CI of correlation (0.746, 0.982). Subsequently, to evaluate Rasch model I evaluated predictive performance after cross-validation with Bradley-Terry model, which was significant, $\chi^2(9) = 64.725$, *p* > 0.001, and therefore also suggested poor performance of the predictive power.

Miscellaneous findings

At the end of the questionnaire, respondents were asked to answer 5 questions concerning every scale used. Those questions were as follows:

FB1 It was fun for me to fill out the questionnaire using this scale.

FB2 Filling out this questionnaire was challenging for me. (*)

FB3 It was easy for me to respond to the questionnaire using this scale.

FB4 The way of filling out the questionnaire with this scale was understandable for me.

FB5 I think that with the help of this scale, I was able to answer the questionnaire fairly accurately.

Primary aim of surveying this was to evaluate participants feelings about usage of the pairwise comparisons scales. I conducted Mann-Whitney U analyses for every pair of scores between Likert scale and pairwise comparisons and all came to suggest significant difference in Likert scale scoring higher except on reverse-worded question FB2, where Likert scale scored significantly lower than pairwise comparisons. While this is not surprising, as Likert scale is well known and expected in questionnaires by respondents.

Table 4: Mean visual graphic scale agreements with statements

	FB1	FB2	FB3	FB4	FB5
Likert scale	71.7	26.4	73.3	85.1	70.0
Pairwise comparisons	66.3	37.1	63.9	80.3	62.5

More interesting is that it not fairly distant from Likert scale in liking even when pairwise comparisons took on average 2.64 times more time to answer to, than Likert scale (this is with 25 pairwise comparisons, *M(pairwise)* = 224.85, *M(Likert)* = 85.32). Participants had also opportunity to comment on anything from the questionnaire and in some cases, they expressed their feelings about the scale used. Out of this comment section, comments concerning pairwise comparisons predominantly expressed difficulty with selection of the items when they both were located on either low side of the scale or high side of the scale (e.g. items v6 and v8). Further concerns included repetition of the same combinations, although this was only perceived, as participants get acquainted with the items as the algorithm provided did not allow any repetition, then it was repetition of the same statements as such, and cognitive overload. On the other side, respondents reported that pairwise comparisons were new to them, and they enjoyed this new scale, and/or usage of intuition while responding to it. There were only 2 comments addressed for the Likert scale, and they both were concerning the restrictive nature of having just 4 levels to choose from. As memorial I chose these next two comments that could represent main thoughts about the scarcely used response scale: *"Very peculiar questionnaire, sometimes completely illogical... But it was at least something new. But what you want to find out from this, I can't imagine. Unless it's about whether such a type of questionnaire can even work. Or maybe the height of your respondents. Well, whatever..."* as a side note, I really learned the height of my sample. *"I trust that you will find what you need. Whether it will be through this methodology, I'm not sure. All the best."* And in line with the old psychological cliché: "If you want to understand person's actions, maybe just ask him directly about them," this could be one of the biggest contributions in evaluating respondent's perception of the used scale.

Discussion

The aim of this thesis was to support realistic view of the measurement in psychology and existence of quantitative latent train trait which underlies item responses (Borsboom et al., 2003; Lord et al., 1968). This was elaborated via Suppes' validation of measurement concept by being able to estimate mathematical ratio of some quantitative variable using two (or more) different scales (Suppes, 1969). The thesis primary concern was therefore the estimation of item parameters using Bradley-Terry model and Rasch model and their subsequent comparison. As a side measures, it also aimed at raising awareness of the need for well-designed and openly reported research. It strived to emphasize the importance of the measurement theory used, checking its assumptions, reliability, and validity (F. M. Lord et al., 1968). As only thus well conducted research is really able to deliver results of "appreciable equally by all observers who are not so abnormal as to fail to appreciate their meaning" (Campbell, 1932).

Key findings

To answer the key concer of Karl Popper, whether we are able to conduct experimental empirical research in psychology, I strive to provide full answer (Popper, 1935). The research hypothesis concerning comparison of item parameters estimated using the Bradley-Terry model and the Rasch model, findings were diverse. Regardless of gender, correlations between item parameters from respective models were high, with correlation after attenuation correction for women being r = 0.907 and for men r = 0.952. Independent sample t-test also significantly suggested relationship between the variables with $t(9) = 6.089$, $p > 0.001$ for women and $t(9) = 7.578$, $p > 0.001$ for men. This comes in slightly unintuitive way, as we would generally expect less populated model producing less precise measures. For better portrayal of the comparison of the parameters, it is helpful to look at the respective figures, which included estimated parameters, regression line, 95% confidence interval and 95% prediction interval. Most of the compared item parameters for women fell within the range of confidence interval with only one item parameter of item v9 ("I have enough legroom on the bus."), and all of them fell within the prediction interval. On the other hand, all but one of the men's parameters fell within the confidence interval of regression line, with exception of item v9. In both cases, this item was located proportionately higher on Rasch model logit scale than on Bradley-Terry model logit scale. One of the explanations for this occurrence could be possible involvement of multiple factors in this item, or it's indeterminant nature. This was mentioned as a problem by some participants in the commenting section, with the biggest concern being unclarity if this item aims at standing or sitting position.

However, as it was seen over the Bradley-Terry and Rasch model estimation, two clusters were created, grouping positively and negatively scored items. I suspect that this happened due to the semantic imperfection of the item selection of the shortened version of the Height Inventory, as it rather than just clustering those items also set them far apart on the logit scales. These clusters were significantly different in item difficulties and negatively scored items laid on the lower side of the scale, and the difference between the ending point of negatively scored item' cluster and starting point of positively scored item' cluster was for Rasch model *logit difference* = 1.36 and on Bradley-Terry model *logit difference* = 2.95. This blank space between the clusters of item parameters may and likely does contribute to the elevation of correlation coefficient presented. From the figures plotting comparison of item parameters, it can be seen, that if the clusters were separated and analyzed so, the direction of the relationship in respective clusters would change, although it couldn't be determined whether this change happens in respect to small sample size or real shift in the relationship. It could also be true, that the clusters represent two different dimensions of the HI in this case, and the item v9, which appeared distant from the both clusters on item parameter comparison figures for both sexes, can contain cross-loadings of both such dimensions. This explanation would be moreover plausible, as we see the similar pattern in both sexes independently. Therefore, to explore relationship between the item parameters estimated using different model, further exploration involving sturdier and item-wise better represented inventory is required.

Parameter estimation

In estimation of parameters for Bradley-Terry model, no difference between men and women responses was found (r = 0.99). Responses were therefore used together to estimate combined Bradley-Terry model, which yielded good predictive power R_{PP} = 0.787 (Baguley, 2012). Model assumptions were soundly met and so contributed in good model-fit (Atkinson, 1972; Cox, 1970). In preliminary analysis for Rasch model, differential item functioning was found for both sexes respectively, and so Rasch model analysis was conducted separately with emphasis on women's data, as insufficient number of male respondents was collected (Wright, 1977). From this sample, item parameters were calibrated, although, with poorer model fit. Clustering effect due to the wording of items also contributed more problematic functioning of the model. On the other hand, correlation between person parameters and reported height was high (r = 0.82) and raw variance measure stated that this model accounted for 64.78 % of variation in the data.

Limitation of the models

As was presented and partially commented in the results section, model-fit for Bradley-Terry model was good, in general. Oversufficient observations per each pair made its estimation straightforward and the model as such didn't show any abnormalities within the residuals other. However, just as the comparison of item parameters was affected by their clustering, estimation of Bradley-Terry model suffered also as this made some item comparisons almost deterministic and close-to-none variation could be observed there, which in turn placed a constraint on information entering the model. This can also stand behind lower predictive power of the model and hence some minor revisions of the scale, such as inclusion of the items bridging this gap would be highly recommended.

Rasch model's validity is restricted by violation of some model assumptions and general rules. Firstly, due to the lower male participants' willingness to respond to online surveys as well as unexpected sampling bias of paid ads on social media I did not collect sufficient number of observations for men (Wright, 1977). It is well known that small sample size disproportionately affects item calibration task as well as assessment of dimensionality (Torre & Hong, 2010) and therefore model for men was incapable to be precisely interpreted for model fit, nor estimated indices. There was, however, yet one more problem connected with ambivalence of the presented items, or rather their possible different functioning. From the commenting section, some possible irregularities in the data may be explained. The most voiced problem was connected to v4 and issue of hugging. 3 commenters stated they usually hug children, and, in this regard, they also used the item. Other such unusual responses connected to specific life experience were found; "I am 165cm but I played basketball when young," "Does the bus item signify legroom when seated or standing," or "It never occurred to me that someone would regard my sister as younger, just because she is shorter than me, actually the opposite, even when I am the older one."

Sample of this study was also, however, plagued by the convenience sampling selected as well as inappropriate Meta platform advertising. It is possible that a different sample would present other ideas concerning questionnaire. It could also be true that the proper measurement of height instead of relying on respondents' answer concerning it, would influence some of the analysis. However, due to the same pattern displayed between male and female participants, it would be unlikely to see completely different outcome in different sample from the same population.

Concluding remarks

In conclusion, this thesis showed support for ability of estimation of similar item parameters using different models and underlying by different measurement theories and their respective philosophies. However, due to the mild violations of the Rasch model assumptions and other limitations, with the most serious being the clustering of the item parameters, it is still questionable if we saw similar results from questionnaire with more evenly distributed item difficulties and it therefore remains in to be determined. This thesis, however, can serve as a plentiful resource of available literature and starting point for future research.

Bibliography

Alfaro-Díaz, C., Esandi, N., Pueyo-Garrigues, M., Canga-Armayor, N., Forjaz, M. J., Rodriguez-Blazquez, C., & Canga-Armayor, A. (2023). Psychometric Evaluation of the Spanish Families Importance in Nursing Care: Nurses' Attitudes Scale Through Classical Test Theory and Rasch Analysis. *Journal of Family Nursing*, *29*(2), 179–191. https://doi.org/10.1177/10748407221148083

Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika*, *38*(1), 123–140. https://doi.org/10.1007/BF02291180

Atkinson, A. C. (1972). A Test of the Linear Logistic and Bradley-Terry Models. *Biometrika*, *59*(1), 37–42. https://doi.org/10.2307/2334612

Baguley, T. (2012). *Serious stats: A guide to advanced statistics for the behavioral sciences*.

Bickhard, M. H. (2001). The Tragedy of Operationalism. *Theory & Psychology*, *11*(1), 35–44. https://doi.org/10.1177/0959354301111002

Bjorner, J. B. (2019). State of the psychometric methods: Comments on the ISOQOL SIG psychometric papers. *Journal of Patient-Reported Outcomes*, *3*(1), 49. https://doi.org/10.1186/s41687-019-0134-1

Blyth, S. (1994). Karl Pearson and the Correlation Curve. *International Statistical Review / Revue Internationale de Statistique*, *62*(3), 393–403. https://doi.org/10.2307/1403769

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences, 2nd ed* (pp. xvi, 340). Lawrence Erlbaum Associates Publishers.

Boone, W. J. (2016). Rasch Analysis for Instrument Development: Why, When, and How? *CBE Life Sciences Education*, *15*(4), rm4. https://doi.org/10.1187/cbe.16-04-0148

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch Analysis in the Human Sciences*. Springer Netherlands. https://doi.org/10.1007/978-94-007-6857-4

Boring, E. G. (1961). The Beginning and Growth of Measurement in Psychology. *Isis*, *52*(2), 238–257.

Borsboom, D. (2005). *Measuring the Mind: Conceptual Issues in Contemporary Psychometrics* (pp. viii, 185). Cambridge University Press. https://doi.org/10.1017/CBO9780511490026

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review*, *110*(2), 203–219. https://doi.org/10.1037/0033-295X.110.2.203

Bradley, R. A. (1976). A Biometrics Invited Paper. Science, Statistics, and Paired Comparisons. *Biometrics*, *32*(2), 213–239. https://doi.org/10.2307/2529494

Bradley, R. A., & Terry, M. E. (1952). Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons. *Biometrika*, *39*(3/4), 324–345. https://doi.org/10.2307/2334029

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

Buckingham, B. R. (1921). Mathematical Ability as Related to General Intelligence. *School Science and Mathematics*, *21*(3), 205–215. https://doi.org/10.1111/j.1949-8594.1921.tb07957.x

Bulmer, M. (2003). Francis galton: Pioneer of heredity and biometry. *Francis Galton: Pioneer of Heredity and Biometry*, 1–357.

Campbell, N. R. (1920). *Physics, the elements*. Cambridge, University Press. http://archive.org/details/physicselements00campuoft

Campbell, N. R. (1932). Quantitative Estimates of Sensory Events. *Nature*, *130*(3291), 809–810. https://doi.org/10.1038/130809b0

Campbell, N. R. (1940). Final report of a committee appointed by the British Association for the Advancement of Science in 1932 to consider the possibility of measuring intensities of human sensation: A commentary. *Advancement of Science*.

Caron, F., & Doucet, A. (2012). Efficient Bayesian Inference for Generalized Bradley–Terry Models. *Journal of Computational and Graphical Statistics*, *21*(1), 174–196. https://doi.org/10.1080/10618600.2012.638220

Causey, R. L. (1969). Derived Measurement, Dimensions, and Dimensional Analysis. *Philosophy of Science*, *36*(3), 252–270.

Chakravartty, A. (2017). Scientific Realism. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2017). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2017/entries/scientific-realism/

Chang, H. (2021). Operationalism. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2021). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2021/entries/operationalism/

Cígler, H. (2019). *Brněnský dotazník výšky (zkrácená verze)*. http://fssvm6.fss.muni.cz/vyska/

Cígler, H., & Palíšek, P. (2023, March 13). *Validita diagnostického nástroje* [Power Point Presentation].

Cox, D. R. (1970). *The analysis of binary data*. Methuen.

Dummett, M. (1982). Realism. *Synthese*, *52*(1), 55–112.

Dunn, P. K., & Smyth, G. K. (2018). *Generalized Linear Models With Examples in R*. Springer. https://doi.org/10.1007/978-1-4419-0118-7

Embretson, S. E., & Reise, S. P. (2000). *Item Response Theory for Psychologists*. Psychology Press. https://doi.org/10.4324/9781410605269

Everett, D. (2010). Explorations in statistics: Correlation. *Advances in Physiology Education*, *34*, 186–191. https://doi.org/10.1152/advan.00068.2010

Fan, Z. (2018, November 28). *Lecture 24—The Bradley-Terry model* [Lecture notes]. https://web.stanford.edu/class/archive/stats/stats200/stats200.1172/Lecture24.pdf

Ferguson, A. (1932). Quantitative Estimates of Sensory Events. *Nature*, *130*(3291), 810–810. https://doi.org/10.1038/130810a0

Ferguson, A., Myers, C. S., Bartlett, R. J., Banister, H., Bartlett, F. C., Brown, W., Campbell, N. R., Craik, K. J. W., Drever, J., Guild, J., Houstoun, R. A., Irwin, J. O., Kaye, G. W. C., Philpott, S. J. F., Richardson, L. F., Shaxby, J. H., Smith, T., Thouless, R. H., & Tucker, W. S. (1940). Quantitative Estimates of Sensory Events: Final report of the committee appointed to consider and report upon the possibility of quantitative estimates of sensory events. *Advancement of Science*, *1*, 331–349.

Ford, L. R. (1957). Solution of a Ranking Problem from Binary Comparisons. *The American Mathematical Monthly*, *64*(8), 28–33. https://doi.org/10.2307/2308513

Galton, F. (1883). *Inquiries into human faculty and its development* (pp. xii, 387). MacMillan Co. https://doi.org/10.1037/14178-000

Galton, F. (1886). Regression Towards Mediocrity in Hereditary Stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, *15*, 246–263. https://doi.org/10.2307/2841583

Galton, F. (1908). *Memories of my life* (2d ed). Methuen & Co.

Galton, F. (1997). I. Co-relations and their measurement, chiefly from anthropometric data. *Proceedings of the Royal Society of London*, *45*(273–279), 135–145. https://doi.org/10.1098/rspl.1888.0082

Galton, Francis, Sir. (1889). *Natural inheritance*. London, Macmillan. https://www.biodiversitylibrary.org/item/76404

Glickman, M. E. (2013). Introductory note to 1928 (= 1929). In E. Zermelo, H.-D. Ebbinghaus, & A. Kanamori (Eds.), *Ernst Zermelo—Collected Works/Gesammelte Werke II: Volume II/Band II - Calculus of Variations, Applied Mathematics, and Physics/Variationsrechnung, Angewandte Mathematik und Physik* (pp. 616–671). Springer. https://doi.org/10.1007/978-3-540-70856-8_13

Gorgi, P., Koopman, S., & Lit, R. (2019). The Analysis and Forecasting of Tennis Matches by using a High Dimensional Dynamic Model. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *182*. https://doi.org/10.1111/rssa.12464

Gustafsson, J.-E. (1980). Testing and obtaining fit of data to the Rasch model. *British Journal of Mathematical and Statistical Psychology*, *33*(2), 205–233. https://doi.org/10.1111/j.2044-8317.1980.tb00609.x

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Kluwer-Nijhoff Pub. ; Distributors for North America, Kluwer Boston.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory* (pp. x, 174). Sage Publications, Inc.

Hamilton, I., Tawn, N., & Firth, D. (2023). *The many routes to the ubiquitous Bradley-Terry model* (arXiv:2312.13619). arXiv. http://arxiv.org/abs/2312.13619

Harding, S. G. (Ed.). (1976). *Can Theories be Refuted?: Essays on the Duhem-Quine Thesis*. Springer Netherlands. https://doi.org/10.1007/978-94-010-1863-0

Hubatková, P. (2020). *Vztah délky Likertovy škály a jejích psychometrických charakteristik* [Bakalářská práce]. Masarykova univerzita, Fakulta sociálních studií. https://is.muni.cz/th/pc9dq/

Humphry, S., Montuoro, P., & Maxwell, C. (2024). Cumulative Ordering as Evidence of Construct Validity for Assessments of Developmental Attributes. *Journal of Psychoeducational Assessment*, *42*(1), 60–73. https://doi.org/10.1177/07342829231199007

Hunter, D. R. (2004). MM Algorithms for Generalized Bradley-Terry Models. *The Annals of Statistics*, *32*(1), 384–406.

Kanamori, A. (2004). Zermelo and Set Theory. *The Bulletin of Symbolic Logic*, *10*(4), 487–553.

Kuhn, T. S. (1962). *The structure of scientific revolutions*. Chicago.

Lazarsfeld, P. (1950). The logical and mathematical foundation of latent structure analysis. In F. Osborn, L. S. Cottrell, L. C. DeVinney, C. I. Hovland, J. M. Russell, S. A. Stouffer, & P. Lazarsfeld (Eds.), *Measurement and prediction: Vol. IV* (pp. 362–412). Princeton University Press.

Linacre, J. (2002a). Understanding Rasch measurement: Optimizing rating scale category effectiveness. *Journal of Applied Measurement*, *3*, 85–106.

Linacre, J. (2002b). What do infit and outfit, mean-square and standardized mean? *Rasch Meas Trans*, *16*.

Linacre, J. M. (n.d.). *Fit diagnosis: Infit outfit mean-square standardized: Winsteps Help*. Winsteps. Retrieved May 13, 2024, from https://www.winsteps.com/winman/misfitdiagnosis.htm

Linacre, J. M. (2005). Rasch dichotomous model vs. One-parameter Logistic Model. *Rasch Measurement Transactions*, *19*(3), 1032.

Lissitz, R. W., & Samuelsen, K. (2007). A Suggested Change in Terminology and Emphasis Regarding Validity and Education. *Educational Researcher*, *36*(8), 437–448. https://doi.org/10.3102/0013189X07311286

Lord, F. (1952). A theory of test scores. *Psychometric Monographs*, *7*, x, 84–x, 84.

Lord, F. M., Novick, M. R., & Birnbaum, A. (1968). *Statistical theories of mental test scores*. Addison-Wesley.

Luce, R. D., & Tukey, J. W. (1964). Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of Mathematical Psychology*, *1*(1), 1–27. https://doi.org/10.1016/0022-2496(64)90015-X

Mair, P., & Hatzinger, R. (2007). Extended Rasch Modeling: The eRm Package for the Application of IRT Models in R. *Journal of Statistical Software*, *20*, 1–20. https://doi.org/10.18637/jss.v020.i09

Martincová, V. (2024). *Střední bod v Likertově škále: Vliv na psychometrické charakteristiky škály a odpověďové procesy respondentů [online]* [Bakalářská práce, Masarykova univerzita, Fakulta sociálních studií, Brno]. https://is.muni.cz/th/ykqmh/

Matthews, J. N. S., & Morris, K. P. (1995). An Application of Bradley-Terry-Type Models to the Measurement of Pain. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *44*(2), 243–255. https://doi.org/10.2307/2986348

Michell, J. (1997). Quantitative science and the definition of measurement in psychology. *British Journal of Psychology*, *88*(3), 355–383. https://doi.org/10.1111/j.2044-8295.1997.tb02641.x

Michell, J. (2008). Is psychometrics pathological science? *Measurement: Interdisciplinary Research and Perspectives*, *6*(1–2), 7–24. https://doi.org/10.1080/15366360802035489

Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology, 3*(1), 1–18. https://doi.org/10.1016/0022-2496(66)90002-2

Parker Jones, O., Alfaro-Almagro, F., & Jbabdi, S. (2018). An empirical, 21st century evaluation of phrenology. *Cortex*, *106*, 26–35. https://doi.org/10.1016/j.cortex.2018.04.011

Pearson, K., & Filon, L. N. G. (1898). On the probable errors of frequency constants and on the influence of random selection on variation and correlation. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, *191*, 229–311. https://doi.org/10.1098/rsta.1898.0007

Popper, K. R. (1935). *The Logic of Scientific Discovery*. Routledge.

Posit team. (2024). *RStudio: Integrated Development Environment for R*. Posit Software, PBC. http://www.posit.co/

Quittre, V., & Monseur, C. (2010). *EXPLORING LOCAL ITEM DEPENDENCY FOR ITEMS CLUSTERED AROUND COMMON READING PASSAGE IN PIRLS DATA*. https://api.semanticscholar.org/CorpusID:18443246

Rasch, G. (1960). *Probabilistic model for some intelligence and achievement tests.* Danish Institute for Educational Research.

Rečka, K. (2018). *Dotazník výšky a váhy [online]* [Diplomová práce, Masarykova univerzita, Fakulta sociálních studií, Brno]. https://is.muni.cz/th/ug7c2/

Reese, T. W. (1943). The application of the theory of physical measurement to the measurement of psychological magnitudes, with three experimental examples. *Psychological Monographs, 55*(3), i–89. https://doi.org/10.1037/h0093539

Rindskopf, D. (2001). Reliability: Measurement. In N. J. Smelser & P. B. Baltes (Eds.), *International Encyclopedia of the Social & Behavioral Sciences* (pp. 13023–13028). Pergamon. https://doi.org/10.1016/B0-08-043076-7/00722-1

Robitzsch, A., Kiefer, T., & Wu, M. (2024). *TAM: Test Analysis Modules*. https://CRAN.R-project.org/package=TAM

Salzberger, T. (2003). Information: When gaps can be bridged. *Rasch MeasurementTransactions*, *17*(1), 910–911.

Savalei, V. (2006). Logistic Approximation to the Normal: The KL Rationale. *Psychometrika*, *71*(4), 763–767. https://doi.org/10.1007/s11336-004-1237-y

Smith, W. (2008). Does Gender Influence Online Survey Participation? A Record-Linkage Analysis of University Faculty Online Survey Response Behavior. *Online Submission*.

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology*, *3*, 271–295.

Stevens, S. S. (1936). A scale for the measurement of a psychological magnitude: Loudness. *Psychological Review*, *43*(5), 405–416. https://doi.org/10.1037/h0058773

Stevens, S. S. (1946). On the Theory of Scales of Measurement. *Science*, *103*(2684), 677–680.

Steyer, R. (2001). Classical (Psychometric) Test Theory. In N. J. Smelser & P. B. Baltes (Eds.), *International Encyclopedia of the Social & Behavioral Sciences* (pp. 1955–1962). Pergamon. https://doi.org/10.1016/B0-08-043076-7/00721-X

Stigler, S. M. (1989). Francis Galton’s Account of the Invention of Correlation. *Statistical Science*, *4*(2), 73–79. https://doi.org/10.1214/ss/1177012580

Suppes, P. (1969). Measurement, Empirical Meaningfulness, and Three-Valued Logic. In P. Suppes (Ed.), *Studies in the Methodology and Foundations of Science: Selected*

Papers from 1951 to 1969 (pp. 65–80). Springer Netherlands. https://doi.org/10.1007/978-94-017-3173-7_5

Suppes, P., & Zinnes, J. (1963). Basic Measurement Theory. In D. Luce & R. Bush (Eds.), *Handbook of mathematical psychology, Volume I* (1–2). John Wiley & Sons.

Tancoš, M. (2019). *Vliv verbálních kotev Likertovy škály na psychometrické charakteristiky dotazníků [online]* [Bakalářská práce, Masarykova univerzita, Fakulta sociálních studií, Brno]. https://is.muni.cz/th/uk8cb/

Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology, 33*, 529–554. https://doi.org/10.1086/214483

Thurstone, L. L. (1954). The measurement of values. *Psychological Review, 61*(1), 47–58. https://doi.org/10.1037/h0060035

Toomela, A. (2007). Culture of science: Strange history of the methodological thinking in psychology. *Integrative Psychological & Behavioral Science, 41*(1), 6–20. https://doi.org/10.1007/s12124-007-9004-0

Torre, J. de la, & Hong, Y. (2010). Parameter Estimation With Small Sample Size A Higher-Order IRT Model Approach. *Applied Psychological Measurement, 34*(4), 267–285. https://doi.org/10.1177/0146621608329501

Trafimow, D. (2016). The attenuation of correlation coefficients: A statistical literacy issue. *Teaching Statistics, 38*(1), 25–28. https://doi.org/10.1111/test.12087

Traub, R. E. (1997). Classical Test Theory in Historical Perspective. *Educational Measurement: Issues and Practice, 16*(4), 8–14.

van der Linden, W. J. (2010). Item Response Theory. In P. Peterson, E. Baker, & B. McGaw (Eds.), *International Encyclopedia of Education (Third Edition)* (pp. 81–88). Elsevier. https://doi.org/10.1016/B978-0-08-044894-7.00250-5

Ward, L. M. (2017). S. S. Stevens's Invariant Legacy: Scale Types and the Power Law. *The American Journal of Psychology*, *130*(4), 401–412. https://doi.org/10.5406/amerjpsyc.130.4.0401

Whelan, J. T., & Klein, J. E. (2022). Bradley-Terry Modeling with Multiple Game Outcomes with Applications to College Hockey. *Mathematics for Application*, *10*(2), 157–177. https://doi.org/10.13164/ma.2021.13

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Wickham, H., François, R., Henry, L., Müller, K., & Vaughan, D. (2023). *dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr

William Revelle. (2024). *psych: Procedures for Psychological, Psychometric, and Personality Research*. Northwestern University. https://CRAN.R-project.org/package=psych

Wright, B. D. (1977). Solving Measurement Problems with the Rasch Model. *Journal of Educational Measurement*, *14*(2), 97–116.

Wright, B. D., & Masters, G. N. (1990). Computation of OUTFIT and INFIT Statistics. *Rasch Measurement Transactions*, *3*(4).

Wright, B., & Panchapakesan, N. (1969). A Procedure for Sample-Free Item Analysis. *Educational and Psychological Measurement*, *29*(1), 23–48. https://doi.org/10.1177/001316446902900102

Wu, W., Junker, B. W., & Niezink, N. M. D. (2022, May 9). *Asymptotic comparison of identifying constraints for Bradley-Terry models*. arXiv.Org. https://arxiv.org/abs/2205.04341v1

Wu, W., Niezink, N., & Junker, B. (2022). A Diagnostic Framework for the Bradley–Terry Model. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *185*(Supplement_2), S461–S484. https://doi.org/10.1111/rssa.12959

Zermelo, E. (1929). Die Berechnung der Turnier-Ergebnisse als ein Maximumproblem der Wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift*, *29*(1), 436–460. https://doi.org/10.1007/BF01180541

Zheng, B., & Agresti, A. (2000). Summarizing the predictive power of a generalized linear model. *Statistics in Medicine*, *19*(13), 1771–1781. https://doi.org/10.1002/1097-0258(20000715)19:13<1771::aid-sim485>3.0.co;2-p

Appendix A Complete wording of items used in the measurement

Note: Reversed items are marked by an asterisk (*)

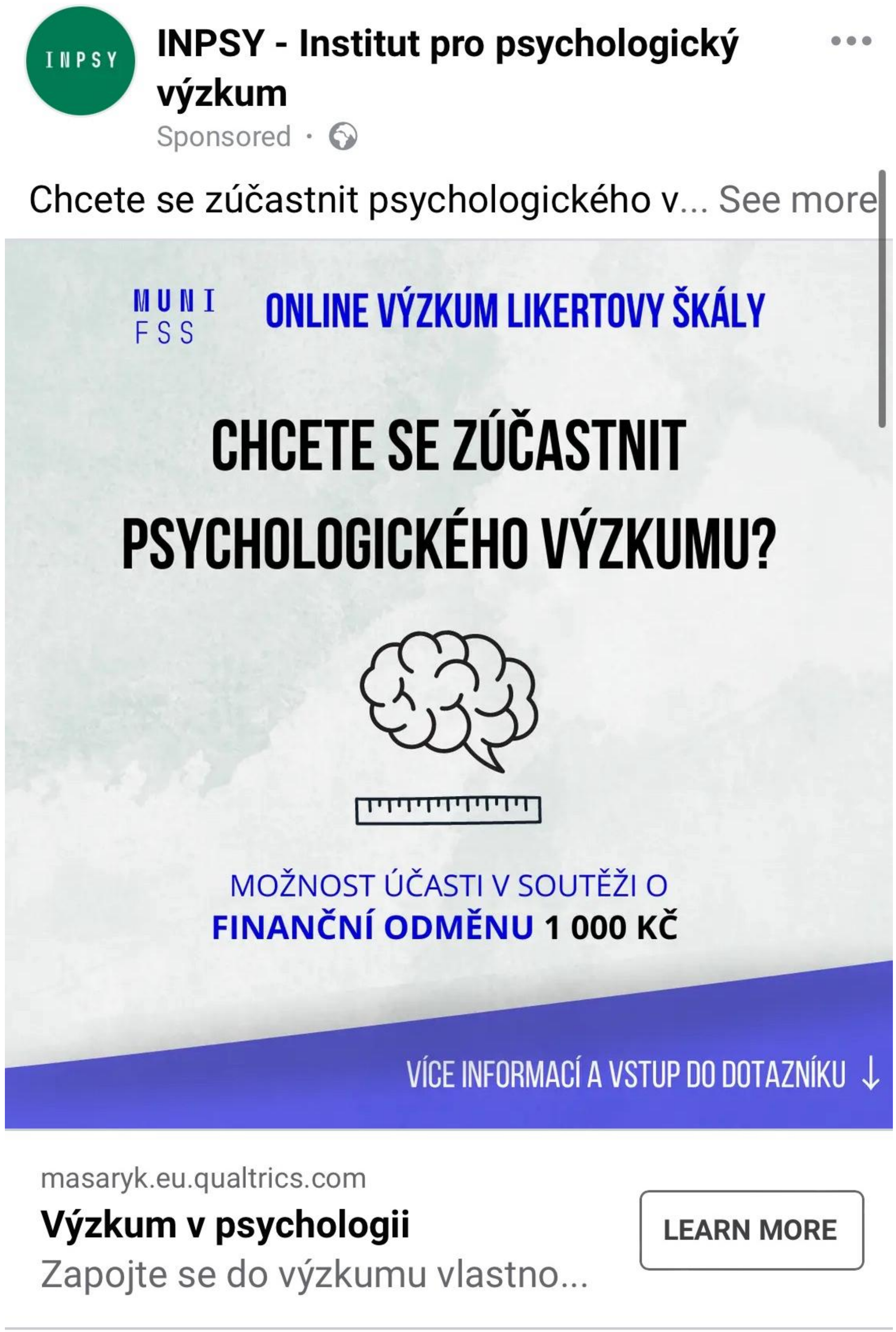
Height Inventory from *Dotazníku výšky a váhy* (Rečka, 2018); shorter version adopted from Tancoš (2018)

V1: „Mám vhodnou výšku na hraní basketbalu nebo volejbalu."
V2: „Slýchávám narážky na to, že jsem vysoký/á."
V3: „Lidem, kteří na koncertě stojí za mnou, většinou má postava dost brání ve výhledu."
V4: „Když chci někoho obejmout, většinou se musím sklonit."
V5: „Často si musím dávat pozor, abych se neuhodil/a hlavou např. o nízký strop nebo rám dveří."
V6: „Často potřebuji stoličku, abych dosáhl/a na něco, na co jiní lidé dosáhnou normálně." (*)
V7: „Jednou z prvních věcí, které si na mně lidé všimnou, je to, jak moc jsem malý/á." (*)
V8: „Často musím stát na špičkách, abych lépe viděl/a." (*)
V9: „V autobuse mívám dostatek prostoru pro nohy." (*)
V10: „Kvůli mé menší výšce lidé hádají, že jsem mladší, než ve skutečnosti jsem." (*)
V11: „Když mluvím s jinými dospělými a chci se jim dívat do očí, častěji na ně spíš vzhlížím nahoru." (*)

English translation

V1: "I have a suitable height for playing basketball or volleyball."
V2: "I often hear remarks about being tall."
V3: "People standing behind me at concerts usually find my stature obstructs their view."
V4: "When I want to hug someone, I usually have to bend down."
V5: "I often have to be careful not to hit my head on low ceilings or door frames."
V6: "I often need a stool to reach something that other people can reach normally." (*)
V7: "One of the first things people notice about me is how small I am." (*)
V8: "I often have to stand on tiptoes to see better." (*)
V9: "I have enough legroom on the bus." (*)
V10: "Because of my shorter height, people guess that I am younger than I actually am." (*)
V11: "When I talk to other adults and want to look them in the eye, I often find myself looking up at them instead." (*)

Appendix B Paid advertisement



INPSY - Institut pro psychologický výzkum

Sponsored ·

Chcete se zúčastnit psychologického v... See more

INPSY
MUNI FSS
ONLINE VÝZKUM LIKERTOVY ŠKÁLY
CHCETE SE ZÚČASTNIT PSYCHOLOGICKÉHO VÝZKUMU?
MOŽNOST ÚČASTI V SOUTĚŽI O FINANČNÍ ODMĚNU 1 000 KČ
VÍCE INFORMACÍ A VSTUP DO DOTAZNÍKU ↓

masaryk.eu.qualtrics.com

Výzkum v psychologii

Zapojte se do výzkumu vlastno...

Appendix C Informed consent, consent with the collection of information and explanation of the study purpose

Děkujeme za váš zájem zúčastnit se výzkumu, který je součástí širšího projektu Škálování, modely měření a odpověďová zkreslení v psychologii, který probíhá v letech 2023–2025 na Masarykově univerzitě. Hlavním řešitelem je Mgr. Hynek Cígler, Ph.D., a tuto studii realizuje Matej Rusiňák. Více informací o projektu **naleznete zde.**

Jaké je téma projektu?
Projekt se zaměřuje na **způsob měření v psychologii**, konkrétně na to, jakým způsobem ovlivňuje formát a znění dotazníku získané odpovědi. Ukazuje se totiž, že různé způsoby položení otázek vedou k různým odpovědím a rozdílné platnosti vědeckých výsledků. Naším cílem je tyto efekty prozkoumat a přispět tak ke zkvalitnění výzkumu v psychologii a sociálních vědách obecně.

Čeho se týká a jak bude probíhat tato konkrétní studie?
V této konkrétní studii se zaměřujeme na samotnou podstatu toho, co psychologickými metodami měříme, a **srovnáváme odpovědi získané pomocí různých metod dotazování.** Kromě tradiční formy, která se vás ptá na míru souhlasu s jednotlivými tvrzeními, budete mít za úkol také srovnávat dvojice tvrzení a vybírat takové, které lépe odpovídá zadání. Podrobné instrukce se dozvíte vždy v příslušné části dotazníku. Celá studie bude trvat asi 10 minut. Dotazník obsahuje i volitelnou, zhruba 5minutovou část, jejímž absolvováním si zdvojnásobíte pravděpodobnost výhry. Poté (budete-li chtít) můžete pokračovat v podobné studii na příbuzné téma (čímž dále zvýšíte pravděpodobnost své výhry).

Jak bude výzkum probíhat?
Na dalších stránkách vás budou čekat konkrétní pokyny a série otázek, týkající se tělesné výšky a základních demografických charakteristik. Žádná z otázek není povinná a kteroukoli z nich můžete přeskočit, v některých částech dotazníku se vám však tlačítko po přeskočení zobrazí až s časovým odstupem. Nejsme si vědomi žádných rizik, které se s účastí ve výzkumu pojí. **Účast ve výzkumu je plně dobrovolná a lze ji kdykoliv předčasně ukončit.**

Jaká je odměna za účast?

Účast ve výzkumu není honorovaná, po ukončení této studie (zhruba v polovině dubna 2024) však budou **vylosováni 3 účastníci**, každý z nich obdrží **finanční odměnu ve výši 1.000 Kč**. Na výhru není právní nárok, ze slosování rovněž budou vyřazeni respondenti, kteří poskytli zjevně nevalidní nebo výrazně neúplná data. Vyplněním volitelné části studie se pravděpodobnost výhry zdvojnásobí. Podmínkou získání výhry je uvedení e-mailové adresy, abychom vás mohli kontaktovat.

Pokud budete souhlasit, na e-mail vám rovněž můžeme poslat pozvánku k zapojení do jiné studie v rámci tohoto nebo navazujícího výzkumného projektu. E-mailová adresa pak může být využita k provázání těchto informací napříč dílčími studiemi, pokud ji v budoucnu rovněž uvedete, nikoli však napříč studiemi z různých projektů. Souhlas s tímto oslovením však není podmínkou pro účast v této výzkumné studii, a ani vás jakkoli nezavazuje k účasti v budoucích studiích.

Podrobný informovaný souhlas si můžete zobrazit níže. Pokud máte jakékoli otázky, neváhejte nás kontaktovat na adrese **520017@mail.muni.cz**.

Informace o zpracování osobních údajů

V rámci tohoto projektu budeme zpracovávat následující osobní údaje: váš věk, pohlaví/gender a vzdělání, odpovědi z dotazníků, kontaktní e-mail pro propojení jednotlivých sběrů dat a losování odměny. K vaší e-mailové adrese budou mít přístup výhradně výzkumníci bezprostředně zapojení do řešení projektu. Po dobu jeho trvání bude možné pomocí vaší e-mailové adresy propojit vaše odpovědi napříč dílčími sběry dat. Ihned po ukončení posledního sběru dat však budou data plně anonymizována a další propojení vašich odpovědí s vaší osobou již nebude možné.
Pokud však budete souhlasit (není podmínkou pro účast v této studii), uchováme si vaši e-mailovou adresu (odděleně od vašich dat) za účelem pozvánky do případných navazujících výzkumů. V takovém případě budeme kromě e-mailové adresy evidovat ještě váš věk, pohlaví/gender a vzdělání, a to za účelem efektivnějšího oslovování v budoucích studiích. Tyto informace budou u nás uloženy po dobu nejdéle pěti let od ukončení tohoto projektu (tedy do prosince 2030), a poté budou smazány. Kdykoliv během této doby můžete také požádat o předčasné ukončení vaší účasti ve výzkumu a tedy i odstranění vaší e-mailové adresy z naší databáze.
Pokračováním v tomto dotazníku rovněž souhlasíte s tím, že plně anonymizovaná data bez jakýchkoli vašich osobních údajů mohou být poskytnuta jiným výzkumníkům za jinými výzkumnými účely, a že mohou být i zveřejněna (např. publikací ve vědecké databázi www.osf.io).
Dále pak:

- Máte právo požadovat přístup k osobním údajům týkajícím se vaší osoby, jejich opravu nebo výmaz, popřípadě omezení zpracování, máte právo vznést námitku proti zpracování osobních údajů týkajících se mé osoby;

- máte právo podat stížnost dozorovému orgánu (Úřad na ochranu osobních údajů) v případě, že se domníváte, že zpracování vašich osobních údajů probíhá v rozporu s právními předpisy;
- máte právo tento souhlas se zpracováním osobních údajů kdykoliv odvolat, aniž by vám za to hrozila jakákoliv sankce či znevýhodnění, a to oznámením na elektronickou adresu cigler@fss.muni.cz, odhlášením se z automaticky rozesílaných e-mailů, případně jinou formou na kontaktní údaje pro zpracování osobních údajů.

Informace o výzkumu:

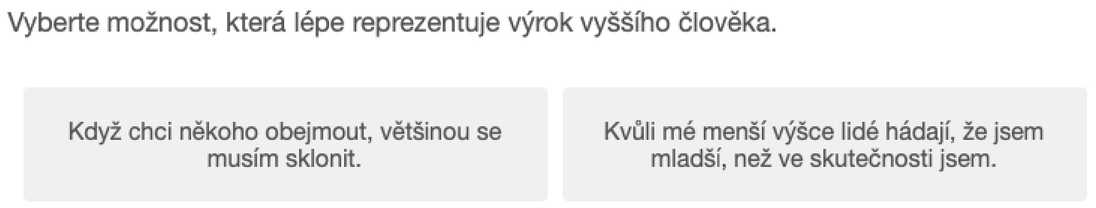
- Název projektu: Vliv formátu odpověďové stupnice na psychometrické parametry položek
- Hlavní výzkumník: Mgr. Hynek Cígler, Ph.D.
- Pracoviště: Fakulta sociálních studií, Masarykova univerzita
- Období řešení projektu: 2023–2025
- Zdroj financování: Grantová agentura České republiky
- V případě jakýchkoli dotazů o tomto výzkumu se můžete obracet na Hynka Cíglera, e-mail cigler@fss.muni.cz

Důležité kontakty:

- Správce osobních údajů: Masarykova univerzita, Žerotínovo nám. 617/9, 601 77 Brno
- Kontaktní osoba správce vašich osobních údajů: Mgr. Hynek Cígler, Ph.D., Joštova 10, 602 00 Brno, cigler@fss.muni.cz, tel. 549 494 616.
- Kontakt na pověřence pro ochranu osobních údajů Masarykovy univerzity: poverenec@muni.cz.
- Tento projekt byl schválen Etickou komisí pro výzkum Masarykovy univerzity. V případě dotazů, nejasností či připomínek k průběhu výzkumu můžete kontaktovat vedení komise na adrese ekv@muni.cz.

Appendix D Example of item pairs for BTM and Pairwise Comparisons

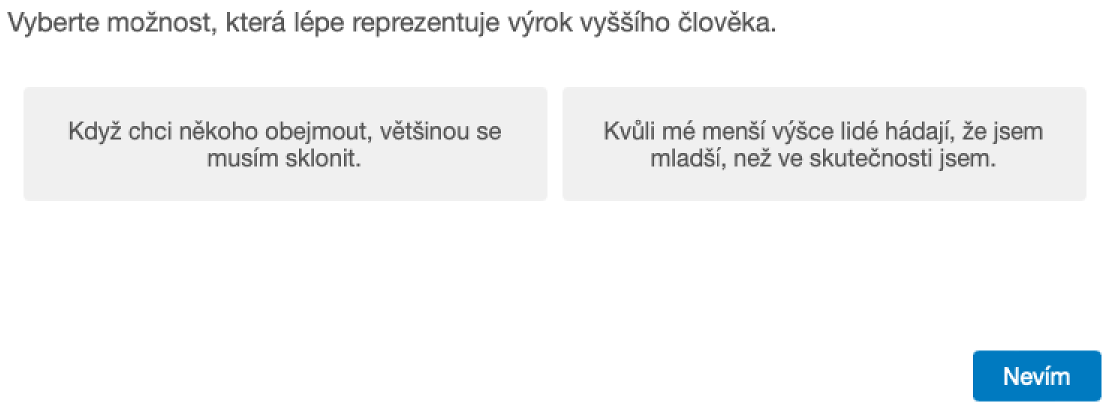
Paired comparison as directly displayed to respondent


Vyberte možnost, která lépe reprezentuje výrok vyššího člověka.

Když chci někoho obejmout, většinou se musím sklonit.

Kvůli mé menší výšce lidé hádají, že jsem mladší, než ve skutečnosti jsem.

Paired comparison 10 seconds after being displayed


Vyberte možnost, která lépe reprezentuje výrok vyššího člověka.

Když chci někoho obejmout, většinou se musím sklonit.

Kvůli mé menší výšce lidé hádají, že jsem mladší, než ve skutečnosti jsem.

Appendix E Selected graphs and figures from RM

Proportion of correct and incorrect responses for women

Items	Responses		Missing
	0	1	
v1	0.625	0.375	0.000
v2	0.791	0.209	0.000
v3	0.773	0.227	0.000
v4	0.849	0.151	0.002
v5	0.862	0.138	0.000
v6*	0.302	0.698	0.002
v7*	0.155	0.845	0.000
v8*	0.432	0.568	0.002
v9*	0.736	0.264	0.002
v10*	0.204	0.796	0.000
v11*	0.455	0.545	0.000
Total proportion	0.562	0.438	0.001

* Reversed items

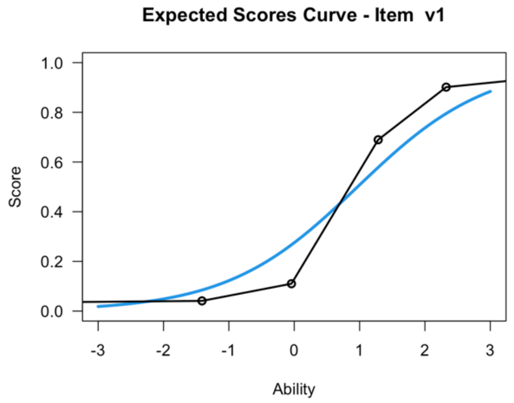
Rasch Model Summary Table for men

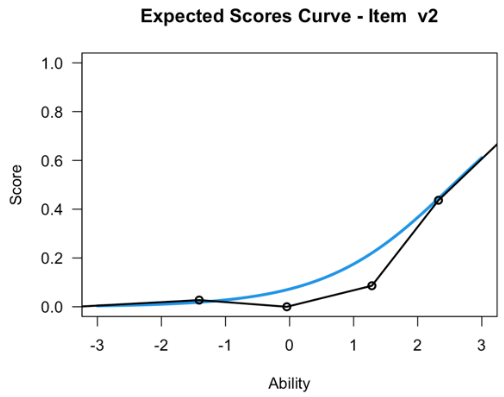
Statistic	Items	Persons
Logit Scale Location Mean	-1.23	0.264
Logit Scale Location SD	2.269	0.443
Standard Error Mean	0.374	0.407
Standard Error SD	0.112	0.494
Outfit MSE Mean	0.933	0.93
Outfit MSE SD	0.863	2.2
Infit MSE Mean	0.879	0.817
Infit MSE SD	0.171	0.847
Std. Outfit Mean	0.091	0.521
Std. Outfit SD	0.679	1.365
Std. Infit Mean	-0.62	-0.32
Std. Infit SD	0.986	1.156
Reliability	0.971	0.775

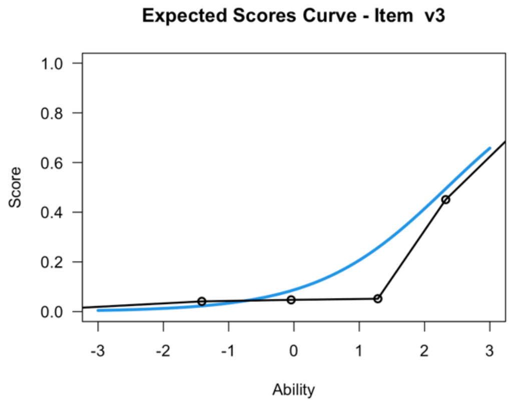
Rasch Model Item Calibration Table for men

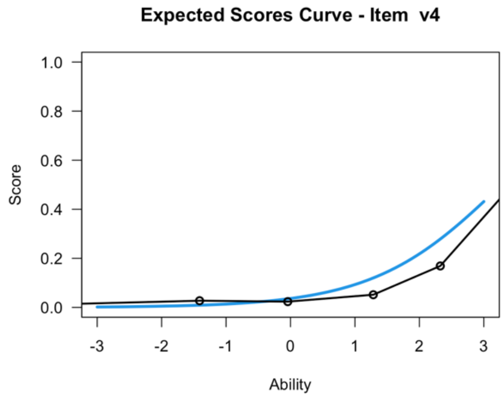
Task ID	Proportion Correct	Item Location	Item SE	Outfit MSE	Std. Outfit	Infit MSE	Std. Infit
v2	0.264	1.716	0.317	0.91	0.265	1.03	0.232
v5	0.341	1.112	0.3	0.533	-0.458	0.636	-2.572
v3	0.407	0.637	0.291	0.657	-0.405	0.873	-0.813
v4	0.429	0.484	0.289	0.858	-0.052	1.091	0.631
v9	0.484	0.11	0.285	0.563	-0.818	0.794	-1.435
v1	0.516	-0.111	0.284	0.662	-0.604	0.817	-1.274
v11	0.857	-2.772	0.382	1.112	0.474	1.091	0.482
v8	0.879	-3.056	0.41	0.5	-0.038	0.741	-1.098
v10	0.879	-3.213	0.427	3.447	1.549	0.895	-0.336
v7	0.923	-3.792	0.498	0.731	0.473	1.064	0.305
v6	0.956	-4.64	0.633	0.292	0.615	0.642	-0.939

Empirical Item Characteristic Curves for women

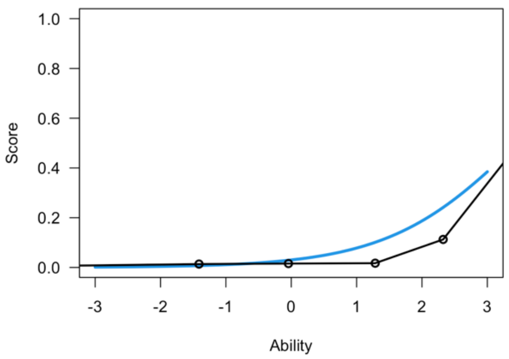




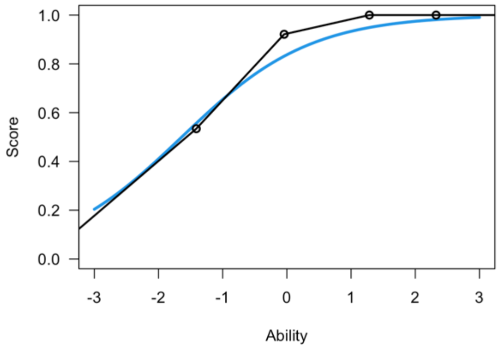




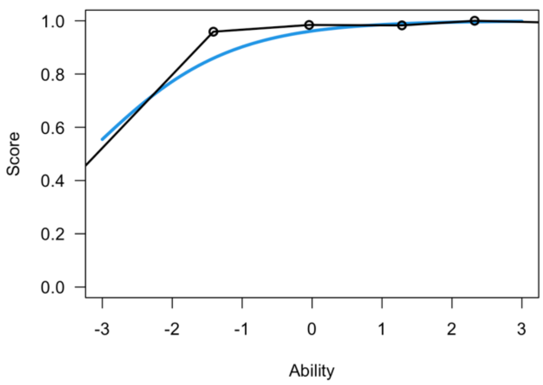
Expected Scores Curve - Item v5



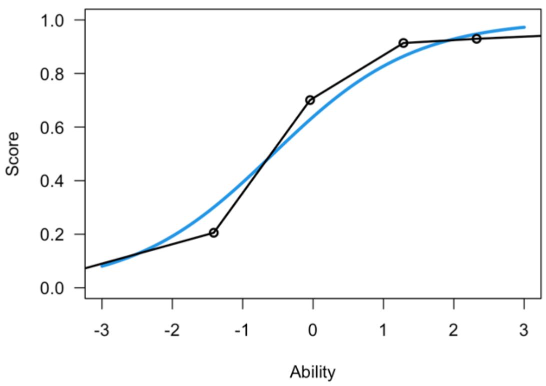
Expected Scores Curve - Item v6

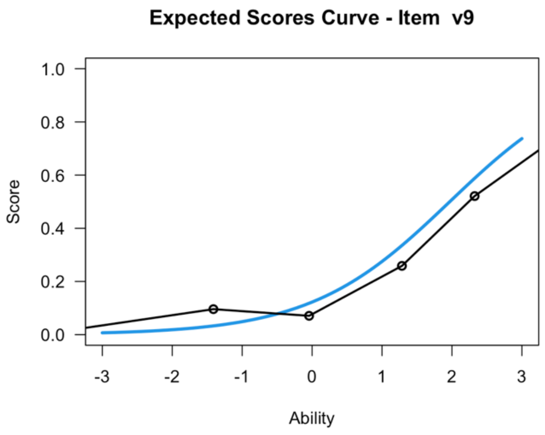


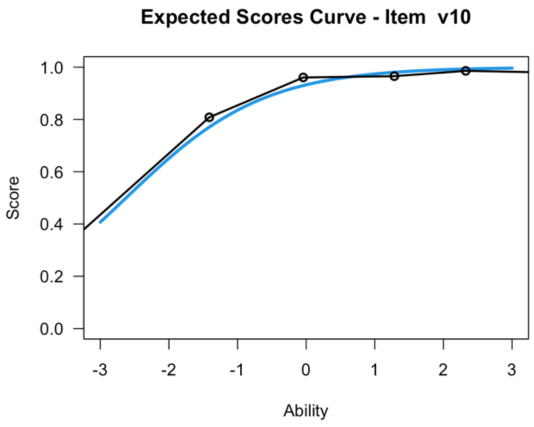
Expected Scores Curve - Item v7

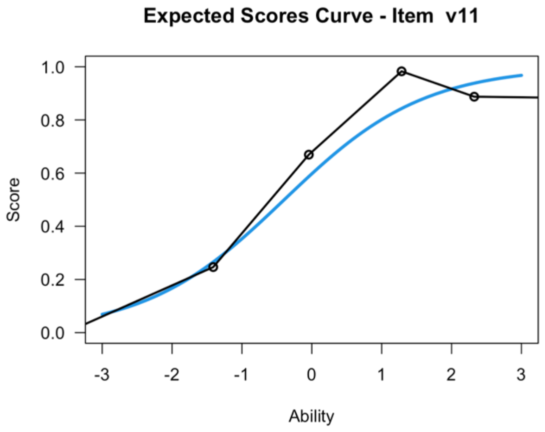


Expected Scores Curve - Item v8




Expected Scores Curve - Item v9
Score
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
Ability


Expected Scores Curve - Item v10
Score
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
Ability


Expected Scores Curve - Item v11
Score
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
Ability