Jørgen Sjaastad has a master's degree in mathematics and a PhD in science education. He is currently a senior researcher at the Nordic Institute for Studies in Innovation, Research and Education (NIFU). His research interests include recruitment to science and mathematics educations and careers, the role played by parents and teachers in young people's attitudes towards science, and psychometrics.

JØRGEN SJAASTAD
Nordic Institute for Studies in Innovation, Research and Education, Norway
jorgen.sjaastad@nifu.no

Enhancing measurement in science and mathematics education research through Rasch analysis: Rationale and properties

Abstract

This article presents the basic rationale of Rasch theory and seven core properties of Rasch modeling; analyses of test targeting, person separation, person fit, item fit, differential item functioning, functioning of response categories and tests of unidimensionality. Illustrative examples are provided consecutively, drawing on Rasch analysis of data from a survey where students in the 9th grade responded to questions regarding their mathematics competence. The relationship between Rasch theory and classical test theory is commented on. Rasch theory provides science and mathematics education researchers with valuable tools to evaluate the psychometric quality of tests and questionnaires and support the development of these.

1. Introduction

1.1 Rasch measurement in science and mathematics education research

The scope of science and mathematics education research includes students' attitudes, interests, opinions, knowledge and skills. Quantitative approaches are often used to investigate these, for instance with regard to attitudes towards science (Bennett & Hogarth, 2009; Sjøberg & Schreiner, 2010), students' evaluation of science and mathematics educations and careers (Oskarsson & Karlsson, 2011), socio-scientific approaches to science and mathematics education (Ekborg, Ottander, Silfver, & Simon, 2012) and students' physics knowledge (Angell, Lie, & Rohatgi, 2011). Turmo and Elstad (2009) describe standardized tests as *"measurement instruments that are used to evaluate levels of specific proficiencies, aptitudes and skills"* (ibid., page 159). Research depends on such instruments to provide data upon which valid inferences can be drawn. However, measuring constructs – *"postulated attribute[s] of people"* (Cronbach & Meehl, 1955, page 283) – may be difficult, and science and mathematics education researchers face many challenges as they engage in human measurement.

Psychometric theory approaches these challenges using mathematically developed models (Raykov & Marcoulides, 2011). Pursuing a model-driven approach to measurement in the 1950's, the Danish mathematician Georg Rasch sought to develop a model satisfying fundamental principles of measurement (Rasch, 1960). The result of his work, known as the Rasch model, is now well established, and applications of the model in science and mathematics education research include studies of attitudes towards science (Zain, Samsudin, Rohandi, & Jusoh, 2010), students' understanding of nature of science (Neumann, Neumann, & Nehm, 2010), mathematics teachers' opinions on teaching practices (Grimbeek & Nisbet, 2006), the influence of science role models (Sjaastad, 2012), and the choice of a tertiary physics education (Oon & Subramaniam, 2012). Moreover, Rasch analysis is fundamental in development of educational assessments like *Programme for International Student Assessment* (PISA), *Trends in International Mathematics and Science Study* (TIMSS) and other international comparative tests (Schulz & Fraillon, 2011; Stubbe, 2011; Wendt, Bos, & Goy, 2011).

Obviously, Rasch analysis is not the only way to conduct psychometric evaluation of instruments. The Rasch model may be framed as a one-parameter logistic model (Raykov & Marcoulides, 2011) and belongs to a family of models developed in psychometric theory to evaluate and improve measures. Classical test theory may also be applied in this respect. Still, there are good reasons for science and mathematics education researchers to gain basic insight to Rasch modeling. Firstly, it applies to most instances where constructs are being measured, and is thus a useful tool for science and mathematics education researchers who develop tests and questionnaires. Secondly, the research community requests documentation of instruments' psychometric quality, which is thoroughly provided through Rasch analysis. Thirdly, insight to Rasch theory is of increasing importance as it enables researchers to be critical readers of the increasing number of articles where Rasch analysis is applied. Finally, such insight may develop the critical sense with respect to instrument validation in general, as Rasch theory points to fundamental principles of measurement.

1.2 The aim of this article

The aim of this article is to provide science and mathematics education researchers with basic insight to the rationale and core properties of the Rasch model. This will enable science and mathematics education researchers to identify instances in own research where Rasch analysis is appropriate and to decipher papers where this is applied. For illustrative purposes, the section about core properties will include results from an authentic Rasch analysis of data from a survey regarding mathematics competence.

The presentation will be given for the polytomous version of the model, where responses have more than two possible outcomes. Many tests and questionnaires in science and mathematics education research include items where partial credit is assigned or where Likert scales with 3, 4, 5 or more response categories are applied. The Rasch model and all properties presented here also applies to the dichotomous case (wrong/correct, no/yes, disagree/agree, etc. scored 0 or 1), except for the analysis of response categories. Two versions of the polytomous model exist; the rating scale model and the partial credit model (Andrich, 1978; Wright & Masters, 1982). The latter is applied here, being a more general version of the former.

The calculations associated with the properties presented here are complex and must be conducted using computer software. Available software include RUMM 2030 (Andrich, Lyne, Sheridan, & Luo, 2011), ConQuest2 (Wu, Adams, Wilson, & Haldane, 2007), WinSteps (Linacre, 2012) and Construct Map (Kennedy, Wilson, Draney, Tutunciyan, & Vorp, 2011), where the latter is free of charge. For the interested reader who wants to learn about Rasch modeling beyond the scope of this article, a friendly yet thorough introduction is provided by Bond and Fox (2007). In addition to the introduction to Rasch analysis provided by Georg Rasch (1960), suggested readings include Ryan (1983), Wright and Stone (1979) and Raykov and Marcoulides (2011). The latter reading introduces the Rasch model using a different rationale than Georg Rasch himself applied as he developed the model.

2. Data

An application of the Rasch model in science and mathematics education research will be presented here; the electronic *NIFU survey* given in the appendix. Nordic Institute for Studies in Innovation, Research and Education (NIFU) developed a questionnaire to investigate Norwegian 9th graders' school experiences and outcomes, including self-evaluation of mathematics competence. The Rasch analysis reported here will draw on data from 933 students.

The students were asked to evaluate their skills with regard to 14 specific mathematics tasks, like calculation of fractions, equation solving and application of Pythagoras sentence (see Appendix). The main question was "*How well do you master the following tasks in mathematics?*" and responses were given electronically by ticking off in one of the five response categories "*not at all*", "*not in particular*", "*somehow*", "*well*", and "*very well*", scored from 0 to 4, respectively. The data were then imported to the Rasch computer program RUMM 2030 for analyses.

3. Rationale of the Rasch model

3.1 A note on the Rasch terminology

The aim of Rasch theory is to develop ways to measure *latent variables*; human attributes which include attitudes, opinions, interests, knowledge, traits, skills, proficiencies and aptitudes. Initially, it was applied to knowledge tests, and expressions such as "person ability" and "item difficulty" were established. As Rasch analysis also applies to measurement of attitudes and opinions, terms like ability and difficulty are interpreted in specific ways. Ability refers to "the amount of a property possessed by a person", which in attitude measurement would point to the level of conviction. Difficulty refers to "the amount of a latent variable a person must possess to have a 50 % chance of receiving score on the item". Score on attitude items is received by endorsing a statement, and full endorsement in Rasch terminology is the "correct" answer. Thus, the most difficult items in attitude measurement are the most extreme statements.

3.2 Unidimensional and invariant measurement

Joel Michell defines measurement in all social sciences similar to measurement in physical science: *Measurement is the estimation or discovery of the ratio of some magnitude of a quantitative attribute to a unit of the same attribute* (Michell, 1997, page 358). Indeed, Rasch theorists seek to develop psychometric instruments holding as many properties as possible as do physical measurement instruments. A fundamental principle in Rasch theory is that individuals' and items' estimates depend on the magnitude of one quantity only, namely, the latent variable of interest. This is referred to as "unidimensional measurement". Indeed, a latent variable may have a multidimensional nature, like "mathematics skills" consisting of an algebra dimension, a geometry dimension, etc. These dimensions might be measured separately. However, if one regards the construct "mathematics skills" as an existing unity about which inferences can be made, this latent variable may be measured along a unidimensional scale to enable researchers to interpret and process test scores in meaningful ways. Whether or not a summary measure is tenable, is an empirical question, which may be investigated using the Rasch model.

Furthermore, measures must be invariant (Bond & Fox, 2007). This means that all persons taking a test, regardless of their ability, identify the same items as most easy and the same items as most difficult. Violation of the invariance principle indicates that something else or something in addition to the latent variable interacts with the persons or the items, causing persons with different ability levels to experience the items difficulties differently.

A Rasch analysis returns estimates relative to a unidimensional interval scale (Andrich, 1988). The dimension is the latent variable of interest, and person- and item-specific numerals are estimated, locating the persons and items on this scale. Persons are assigned an ability B according to how much of the latent variable they possess, and items are assigned a difficulty D according to how much of the variable persons need to have 50 % chance of answering the item correctly. Thus, the Rasch model is probabilistic. The person with ability B in Figure 1 has 50 % chance of answering items with difficulty D3 correctly. He is very likely to answer items with difficulty D1 correctly and D5 incorrectly. Items with difficulty D2 are answered correctly more often than incorrectly, while the opposite is true for items with difficulty D4. Most Rasch computer programs rescale the scores in order to make the mean item difficulty equal to 0. The easiest items will have difficulties below zero and the most difficult items will have estimates above zero. Consequently, the most skilled respondents have high ability estimates, and those who do not possess much of the latent variable have low ability estimates.

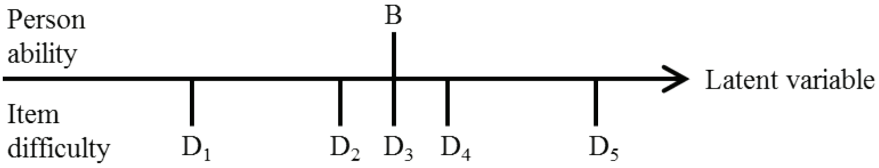


Figure 1. An excerpt from an infinite, unidimensional interval scale representing the amount of a latent variable increasing from left to right. A person's ability B and five item difficulties D1 to D5 are located on the scale.

In instances where the Rasch model is applied to polytomously scored measures, the transitions between two successive score categories are located at the scale. This implies that an item with four score categories (for instance; 0, 1, 2 and 3) will have three points on the logit scale where the probability of receiving the higher scores will become greater than the probability of receiving the lower score. These points are called thresholds, and will be elaborated on in the following section.

3.3 The item characteristic curve (ICC)

The mathematical expression of the dichotomous Rasch model is given in Equation 1. It gives the probability of a person with ability B to answer correctly (score = 1) an item with difficulty D. In line with fundamental properties of measurement, it is apparent in the equation that the probability of correct response depends on the difference between person ability and item difficulty only.

$$P(Score = 1\,|B, D) = \frac{e^{(B-D)}}{1 + e^{(B-D)}} \qquad (1)$$

Notice that with high abilities B, the probability converges towards 1, while high difficulties D make the probability converge towards 0, as *e*-∞ converges towards 0. The probability calculated in Equation 1 is graphically represented by the item characteristic curve (ICC), displaying how the probability of receiving score on an item depends on the difference between person ability and item difficulty. The ICCs of two items, Item 1 with difficulty D1 = -2.0 and Item 2 with difficulty D2 = -0.5, are given in Figure 2. Note that the units on the Rasch scale are called logits (log-units), as the estimates are provided through iterative processes with initial values given by the logarithm of the odds (Andrich & Marias, 2009).

The dashed lines in Figure 2 illustrate how item difficulties are defined. Persons with ability estimates of -2.0 logit have 50 % chance of answering Item 1 correctly, hence the item difficulty of -2.0. Persons with ability -0.5 logits have 50 % chance of receiving score on Item 2, which then has difficulty -0.5. The solid line illustrates that persons with ability -0.5 have an expected value on Item 1 of 0.82, that is, an 82 % chance of answering Item 1 correctly. This value can also be derived by substituting B = -0.5 and D = -2.0 in Equation 1.

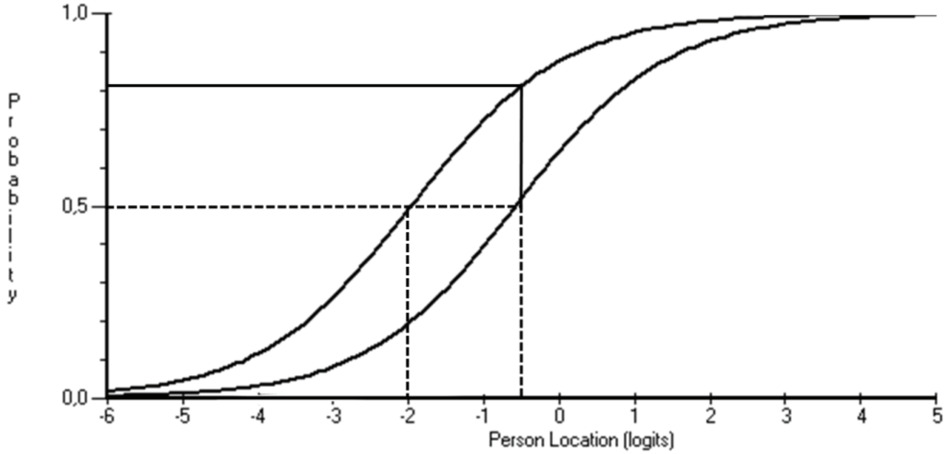


Figure 2. The item characteristic curves (ICC) of two items with difficulties -2.0 (left curve, Item 1) and -0.5 (right curve, Item 2). Illustration from Rasch software RUMM 2030.

At the end of Section 3.2, thresholds were introduced as points on the logit scale where the probability of receiving a specific score is equal to receiving the adjacent higher score. Thus, the thresholds of an item have their own "difficulties" T1, T2, etc. The polytomous Rasch model is expressed by Equation 2, which gives the probability of receiving a score x. M is the number of thresholds (i.e. the item has M+1 response categories) and x thus takes values from 0 to M. Notice that in the dichotomous case, M = 1 and Equation 2 reduces to Equation 1.

$$P(Score = x \mid B, D, T_k) = \frac{e^{(x(B-D)-\sum_{k=1}^{x} T_k}}{\sum_{x=0}^{M} e^{(x(B-D)-\sum_{k=1}^{x} T_k}} \qquad (2)$$

Equation 2 is the mathematical expression of the *rating scale model*. Another polytomous Rasch model exists, namely, the *partial credit model* which will be applied to the NIFU survey data. In the latter, the thresholds are allowed to vary between items in the same instrument. The formula looks identical, but the threshold parameters Tk must be specified for each item in the instrument.

3.4 Items separating and providing information about persons

Items differ in which persons they separate most efficiently. In the Rasch terminology, this is referred to as discrimination. The curves in Figure 2 illustrate in which parts of the logit scale these particular items differentiate most between persons. A small ability increase in the region from -3.0 to -1.0 logit implies a substantial increase in expected value to Item 1, while Item 2 differentiates most between -1.5 logits and 0.5 logits. Consequently, these are the regions where the items provide most information, as they separate efficiently between persons in these intervals. For instance, a person with ability -1.0 logit will get Item 1 correct far more often than a person with ability -3.0. However, Item 1 does not separate as well between two persons with ability 1.0 logit and 3.0 logit, as they have about the same expected value. Both will probably answer the item correctly, and thus the item does not provide information about the difference in ability between the two persons.

The latter example is visualized in Figure 2 by the converging ends of the curves. For instance; a 10 year old knows more mathematics than a 6 year old, but the probability of him solving an equation of second degree has not increased much over the four foregoing years. Similarly, a mathematics professor has higher mathematics ability than an undergraduate student. Still, the probability of the professor solving the equation of second degree has not increased much due to her professional career. They will both have a probability close to 1 of solving the task correctly.

Conclusively, the Rasch model is a probabilistic model that supports the development of invariant measures, locating persons and items on the same unidimensional interval scale. An item's ICC displays how the ability relates to the expected value.

4. Rasch analysis properties

Core properties of Rasch analysis will be presented in Section 4.1 throughout Section 4.7. These properties are important elements of a validity argument, as they account for many of the psychometric qualities of an instrument. Examples from a Rasch analysis conducted on data from the NIFU survey presented in Section 2 will be provided consecutively.

4.1 Is the test well-targeted?

As discussed in Section 3.4, an item provides most information about persons with ability estimates in the interval where the ICC slope is steep (Figure 2), which is centered on the item difficulty estimate. Hence, a well-targeted instrument has most items with difficulties in the same region of the scale as the abilities of the persons taking the test. A few items centered in the lower and higher regions must also be included in order to provide information about persons with low and high ability estimates, respectively.

The items in the NIFU survey are polytomous, which makes the items' thresholds relevant. That is; we want to know at which points along the logit scale a respondent is likely to move from one score category to any of the successive. Having this certain ability estimate, a respondent goes from having higher probability of being in one category to having higher probability of being in one of the successive categories. Items with 5 response categories have 4 thresholds, and the 14 items in the NIFU survey have a total of 56 thresholds. These are displayed in the right part of Figure 3, where the first two digits indicate item number and the third digit indicate item threshold number. For instance, at the bottom right corner is "11.1", located at about -2.20 logit. This is the first threshold of Item 11, meaning that persons with ability estimates below -2.20 are most likely to receive a score of "0", while persons above -2.20 are most likely to receive a score of "1" or higher. The second threshold of Item 11, labelled 11.2, is located at -0.60 logit. Persons above this point are thus more likely to receive a score of 2 or higher than a score of 1.

Rewriting the sentence about well-targeted instruments to apply to the polytomous case, a well-targeted instrument has most thresholds with difficulties in the same region of the scale as the ability of the persons taking the test. The distribution of persons and items along the logit scale is investigated using a person-item map, which in the polytomous case displayed in Figure 3 actually will be a person-threshold map.

Considering the person estimates shown in the left part of Figure 3, the person-item map reveals that the test is slightly easy for these test-takers. Many thresholds are located in the lower regions (between -2.40 and -0.80) where only a few persons have ability estimates. Moreover, the test could benefit from including slightly more difficult items. Above 2.40 logit, where every eight person is located (81+39 out of 933), there are only four thresholds. Many persons are likely to get a full score, which makes it difficult to separate between these persons. This matter is indicated by the PSI, introduced below.

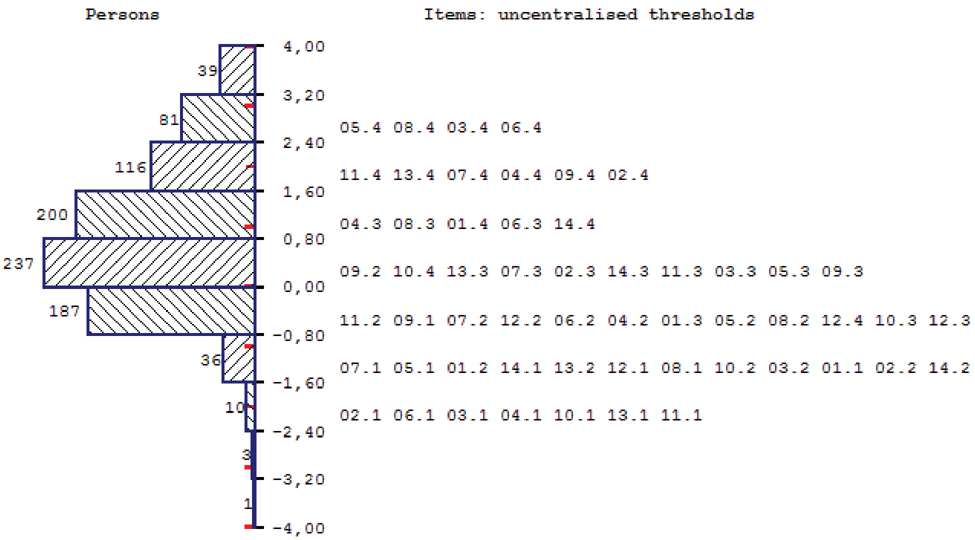


Figure 3. The person-item map of the NIFU study. The vertical line is the unidimensional interval scale representing the amount of mathematics skills. To the left are persons located according to ability estimates and to the right are the 14 items represented by 56 thresholds, located according to difficulty estimates. Illustration from Rasch software RUMM 2030.

4.2 Does the test separate the persons well according to their ability?

Reliability measures provide estimates for the precision of data. In Rasch analysis, statistics are provided indicating how reliably the test separates and ranks the persons. The RUMM software reports a person separation index (PSI) in this respect. This index is comparable to Cronbach's alpha in many ways. The PSI and the alpha both take values between 0 and 1, and the values often turn out quite similar when applied to the same data. The main difference between the two is that the PSI estimates draw on logit values and not on raw scores (Tennant & Conaghan, 2007).

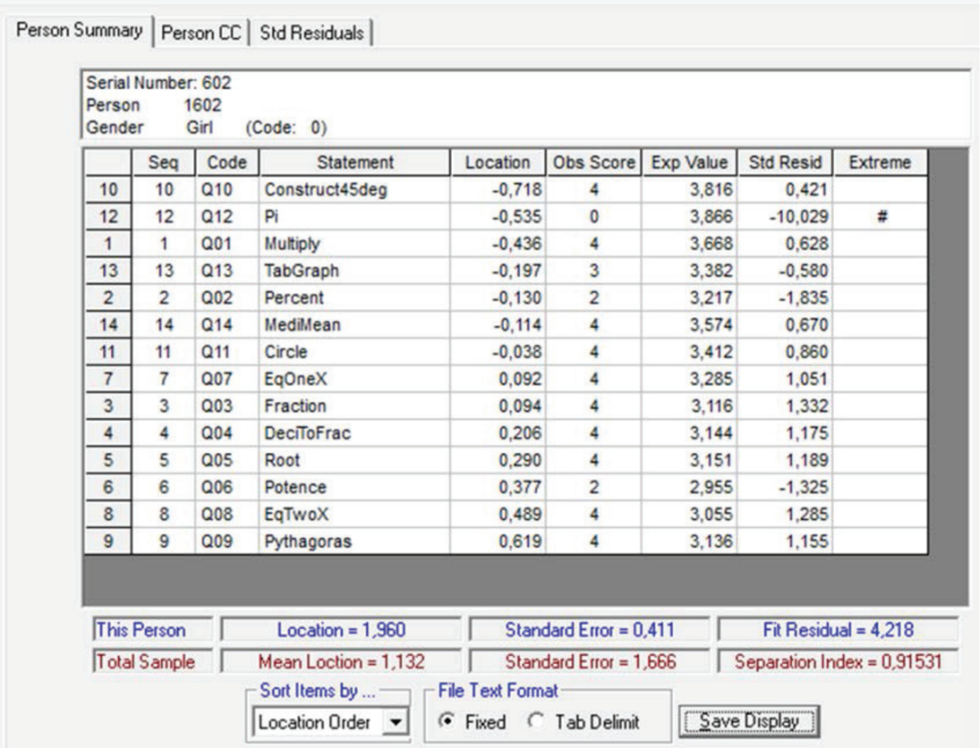
As discussed in Section 3.4, most information about a person is provided by the items with difficulty estimates close to the person's ability. The PSI thus depends on how well-targeted the test is. Moreover, it depends on the ability distribution of the respondents, as it is harder to separate persons when they are close in ability. With respect to the NIFU study, RUMM estimated a PSI of 0.92, which is regarded as good. So even though few thresholds were located in the region of the most able persons (Figure 3), the total number of items in the NIFU study keeps the PSI good. Had the test been more targeted at the most able group of students, the total number of items in the questionnaire could have been reduced and the PSI would remain good.

4.3 How well do the persons' test scores fit the model?

It might seem odd to ask whether or not persons' test scores fit a given model. However, "the rationale is quite straight-forward: Given that all items in a test measure the same latent variable, a respondent is more likely to get easy items correct than difficult items. A Rasch analysis produces individual person fit estimates, indicating whether or not the person's response pattern is as expected. Arranging the items from easy to difficult, an expected response pattern occurs when a respondent gets most items correct up to a certain difficulty D1, and from a certain other difficulty D2 the respondent gets most items incorrect. Between D1 and D2 are the items with difficulties close to the person's ability, where the scoring pattern is more random.

By investigating person fit one might detect cheating, guessing, problems with the instrument or problems with the administration of the instrument. For instance, a person might get all items correct up to a certain level of difficulty, and from this point on get all items incorrect (a perfect "Guttman-pattern"). However, the Rasch model is probabilistic and we expect some variation in the scores received on items close to the person ability, where the probability of correct answer is about 40 %-60 %. So even though the perfect Guttman-pattern is the most likely of all patterns, it is an unlikely pattern relative to the sum of all other patterns. Such a respondent *over-fits* to the Rasch model, calling for further investigation: Did the test consist of only too easy and too hard items, with no item difficulties close to the person ability? A person that *under-fits*, on the other hand, has a response pattern far from the Guttman-pattern, getting many easy items wrong and many difficult items correct. This might indicate lack of concentration or motivation, situational distractions, that the respondent has been guessing or cheating at some hard items, or that the test includes items concerning topics covered in class at a time when the respondent was away from class.

Person fit estimates provided by RUMM have expected values between -2.5 and +2.5, where values close to 0 indicate good fit. Person fit below -2.5 indicate that the person over-fits, while under-fit is indicated by values above +2.5. One of the respondents in the NIFU survey had person a fit estimate of +4.2 and was thus a candidate for further investigation. This student's response pattern is displayed in Figure 4. The items are arranged according to difficulty with the easiest items at the top. According to the observed scores ("Obs Score" column), this student claimed to know all but four tasks "very well". Her scores under-fit due to which four specific mathematics tasks she does not master very well: Three of these are among the five easiest tasks. These scores might result from lack of concentration or interest in the survey.



Serial Number: 602
Person 1602
Gender Girl (Code: 0)

	Seq	Code	Statement	Location	Obs Score	Exp Value	Std Resid	Extreme
10	10	Q10	Construct45deg	-0,718	4	3,816	0,421	
12	12	Q12	Pi	-0,535	0	3,866	-10,029	#
1	1	Q01	Multiply	-0,436	4	3,668	0,628	
13	13	Q13	TabGraph	-0,197	3	3,382	-0,580	
2	2	Q02	Percent	-0,130	2	3,217	-1,835	
14	14	Q14	MediMean	-0,114	4	3,574	0,670	
11	11	Q11	Circle	-0,038	4	3,412	0,860	
7	7	Q07	EqOneX	0,092	4	3,285	1,051	
3	3	Q03	Fraction	0,094	4	3,116	1,332	
4	4	Q04	DeciToFrac	0,206	4	3,144	1,175	
5	5	Q05	Root	0,290	4	3,151	1,189	
6	6	Q06	Potence	0,377	2	2,955	-1,325	
8	8	Q08	EqTwoX	0,489	4	3,055	1,285	
9	9	Q09	Pythagoras	0,619	4	3,136	1,155	

This Person | Location = 1,960 | Standard Error = 0,411 | Fit Residual = 4,218

Total Sample | Mean Loction = 1,132 | Standard Error = 1,666 | Separation Index = 0,91531

Sort Items by ... Location Order

File Text Format ● Fixed ○ Tab Delimit

Figure 4. The scoring pattern of a student with a fit residual of +4.2. Illustration from Rasch software RUMM 2030.

4.4 How well does a specific item fit the other items?

Each item in a test must target the same latent variable as the other items are targeting. Rasch statistics are derived from an analysis of the collection of all items, and thus, an item's statistics indicate whether or not it fits the other items. Following the Rasch terminology, this is called an item's "fit to the model" (Styles & Andrich, 1993). In evaluating an item's fit, all respondents are divided into groups (typically 3-10 groups) according to their ability estimates. All groups' mean scores to the item are calculated. The mean scores of these groups, illustrated by dots in Figure 5, are plotted together with the item characteristic curve. The discrepancy between the observed group values and the theoretically expected values displayed by the ICC indicates how well the item fits the Rasch model. Broadly speaking, such comparisons have three potential outcomes: The item either has good fit, it under-discriminates or it over-discriminates. These three occurrences will be exemplified below.

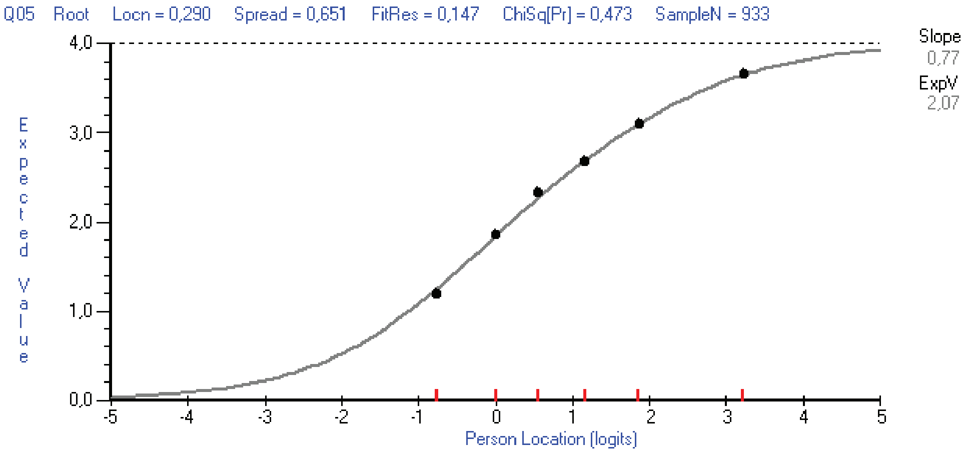


Figure 5. The item characteristic curve (ICC) of Item 5 (item fit 0.1) and observed group mean values. Illustration from Rasch software RUMM 2030.

Items with *good item fit* will have observed values close to the expected values displayed by the ICC. Item 5 in the NIFU questionnaire, displayed in Figure 5, exemplifies this. Respondents in the NIFU study were divided into six groups according to ability. The six observed group mean values on Item 5 out to be located close to the expected values derived from the Rasch model (visualized by the ICC). The fit residual, estimated by taking the standardized sum of the differences between the observed values and the expected values, is reported in the heading of Figure 5 ("FitRes"). As for the person fit estimates, item fit values close to 0 indicate good fit, while values above +2.5 and below -2.5 indicate under-discrimination and over-discrimination, respectively. Item 5 has a fit residual of 0.1 which is very good.

Under-discrimination occurs when the observed values do not increase as much as the increase in ability would suggest. Such items do not separate sufficiently between persons according to the latent variable of interest. The item with the greatest positive fit residual in the NIFU study is displayed in Figure 6. Item 1 has a fit residual +5.7 and under-discriminates. Connecting the observed values (the dots) creates a line which is more horizontal than the ICC, meaning that an increase in ability leads to less increase in observed score than expected by the model. As the persons are assigned abilities according to the latent variable, the immediate inference is that the item at too little extent measures this latent variable.

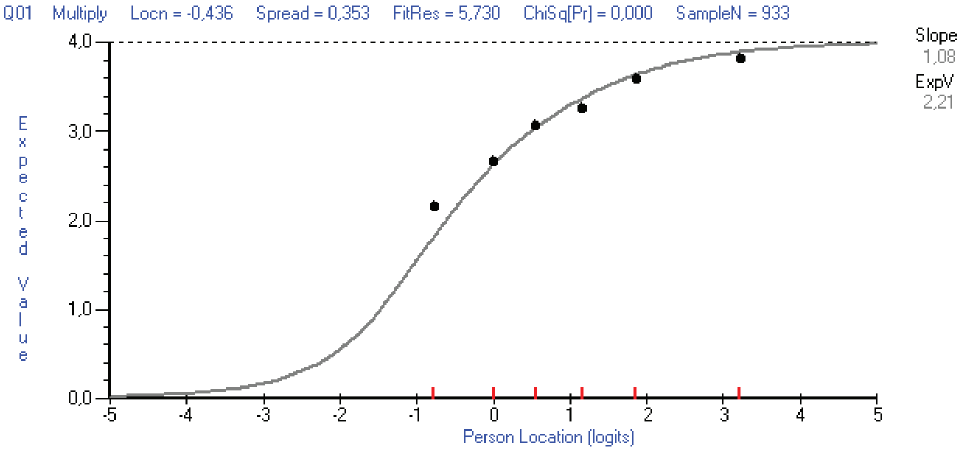


Figure 6. The item characteristic curve (ICC) of Item 1 (item fit +5.7) and observed group mean values. The item under-discriminates slightly. Illustration from Rasch software RUMM 2030.

Over-discrimination means that the item separates persons according to the latent variable, but it does so in a limited region of the logit scale. For the many respondents below or above this interval, most will get the item wrong or correct, respectively, and little information is provided by the item that helps to separate between the persons within these two groups. The curve in Figure 7 displays the ICC of Item 2, having a fit residual -2.9. Connecting the observed values of the six ability groups provides a line more vertical than the ICC, suggesting that the item over-discriminates slightly.

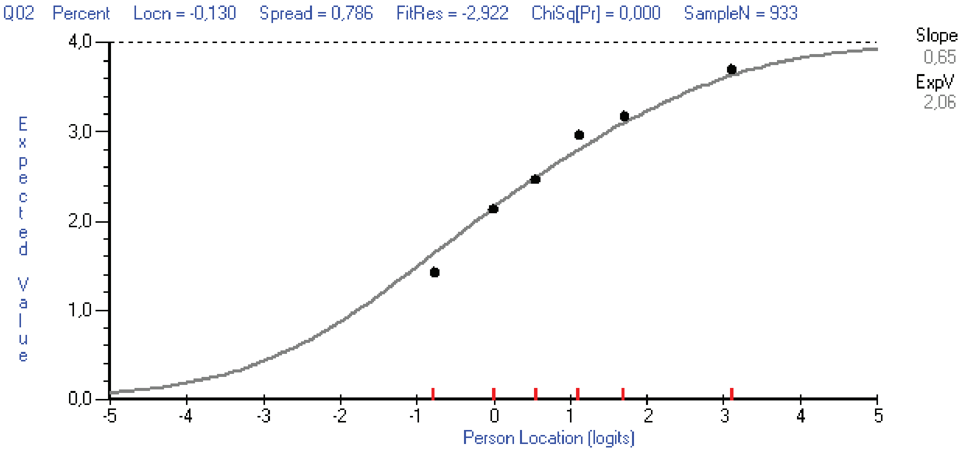


Figure 7. The item characteristic curve (ICC) of Item 2 (item fit -2.9) and observed group mean values. The item over-discriminates slightly. Illustration from Rasch software RUMM 2030.

4.5 Do the items function differently for different groups of persons?

Two persons with the same ability estimate are supposed to have the same expected value on each item in a test, regardless of differences in nationality, gender, occupation, or any latent variable other than the one under investigation (Hagquist & Andrich, 2004). Differential item functioning (DIF) refers to instances where this assumption, and thus the invariance principle presented in Section 3.2, is violated. For instance, DIF with regard to nationality means that the item favors respondents from certain countries. This is discovered by grouping the respondents according to ability estimates, and then calculating the observed mean values for persons from different countries separately within the ability groups. Significant differences in values within the ability group suggest that the item not only measures the latent variable upon which the abilities are estimated, based on all items in the test, but that this item also measures something associated with nationality. An example of such DIF is a physics test including an item on the physics of cross-country skiing. Students in Nordic countries may perform better than expected relatively to other students at this item, not due to their physics competence, but due to contextual understanding and hands-on experiences with the physics of skiing. An example of DIF with regard to gender is provided by Item 10 in the NIFU survey. This is illustrated in Figure 8.

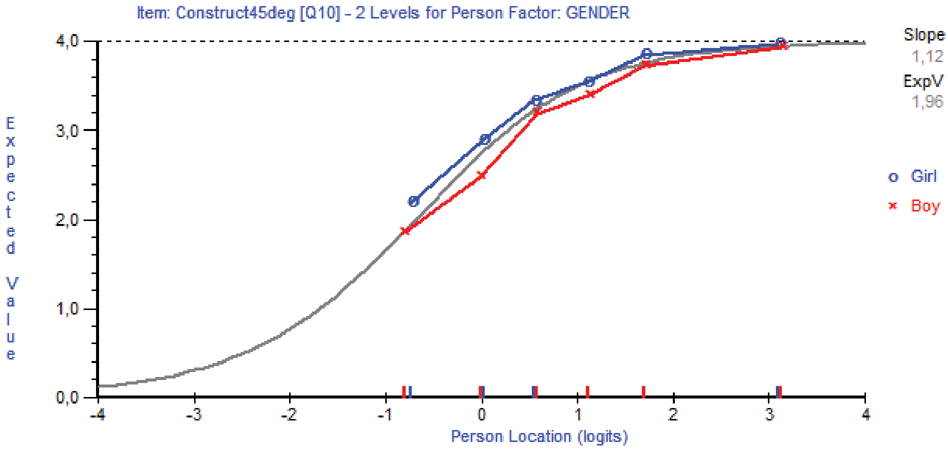


Figure 8. The item characteristic curve (ICC) of Item 10 and observed group mean values for girls (circles) and boys (junctions). The uniform differential item functioning is significant at a 0.05-level, Bonferroni adjusted. Illustration from Rasch software RUMM 2030.

Item 10 regards knowing how to construct a 45 degree angle using math compass and ruler, which is the only "practical geometry" item in this test. From Figure 8 it is apparent that the boys in the NIFU survey have lower observed values on Item 10 than girls with the same ability. The invariance principle is thus violated; the test-takers do not agree on how difficult Item 10 is. Even though the boys may be at the same ability level as the girls, their observed values on this particular item is lower, which means that they would label the item as more difficult than the girls would do. This occurrence of DIF is called uniform DIF, as one group consistently has higher observed values than the other group. Non-uniform DIF occurs when it is the item discrimination that differs between the two groups.

Notably, with regard to the calculation of DIF, it is often advised to let the significance level of group differences undergo a Bonferroni adjustment. RUMM conducts 3 different DIF tests for each item, summing to a total of 42 tests in the case of NIFU's survey. With a 5 % significance level, this means that Type I errors are likely to occur. The Bonferroni adjustment takes the significance level (0.05) and divides it with the number of tests (42), providing an adjusted and more conservative significance level (0.00119). The DIF of Item 10 is still significant, as the probability is 0.00003 (value provided by RUMM) of this difference occurring by chance.

4.6 Does the test produce unidimensional data?

A test satisfying fundamental principles of measurement will produce unidimensional data (Gardner, 1996; Tennant & Pallant, 2006). As mentioned in Section 3.2, the latent variable of interest may have a multidimensional nature. Still, the scale developed to measure the possibly multidimensional latent variable is unidimensional. When persons' expected values (values between 0 and 1) are subtracted from their observed responses (0 or 1), the effect of this dimension is removed and, ideally, only uncorrelated residuals remain. Systematic correlations in the residuals of a subgroup of items would suggest that the items in this subgroup have something in common besides the latent variable of interests, and the data is thus not unidimensional.

Unidimensionality may be investigated by conducting a principal component analysis (PCA) on the person-by-item residual matrix. A person is assigned separate abilities and error estimates relative to the different subgroups of items identified by the PCA. An independent t-test will then reveal if the person scores are significantly different on the different subgroups. Tennant and Conaghan (2007) suggest that if, at a significance level of 5 %, more than 5 % of the respondents do so, the test should be investigated further for multidimensionality. If the latent variable is multidimensional by nature, the PCA is likely to identify these subdimensions. Results from the t-tests must then be combined with other statistics and theoretical considerations in a discussion of whether or not the subdimensions are in fact individual dimensions that should be measured separately.

A PCA of the person-item residuals in the NIFU survey identified a first principal component which accounted for about 17 % of the variance. The three items loading most positively on this component were items 2, 3, and 4, while items 7, 8, and 9 had the most negative loadings. The former items regard calculation of percentages, fractions and decimal numbers, while the three latter items regard solving equations with one or two unknown variables and using Pythagoras' sentence to do calculations on a triangle (see Appendix). One might say that the items in each subgroup have something in common. Figure 9 shows that 11 % of the respondents scored significantly different (0.05-level) on the two subgroups of items. This does not support the argument that the instrument produces unidimensional data. The two subgroups of items assign the respondents with slightly different ability estimates, which is unexpected from tests that allegedly measure the same latent variable.

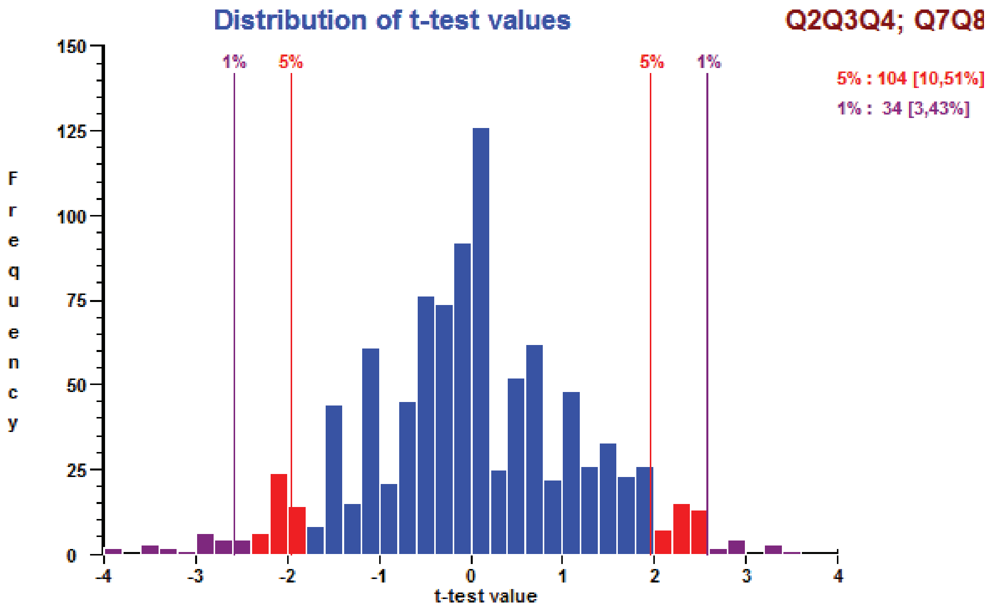


Figure 9. Summary of independent t-tests conducted to investigate differences between respondents' ability estimates on two subsets of items (items 2, 3, and 4 and items 7, 8, and 9). Illustration from Rasch software RUMM 2030.

4.7 Are the response categories appropriate?

Polytomous Rasch models are characterized by items scored in three or more categories. These models apply when persons are assigned partial credit or when responses are given using Likert scales. Ideally, all successive categories represent a natural increase in ability. However, categories may be redundant, as when the natural increase in ability takes the respondent from category 1 to category 3: Scoring a student "2" if he can multiply and "3" if he can multiply and add would probably make category 2 redundant, as most students who know how to multiply also know how to add. With regard to attitude instruments, mid-categories sometimes turn out to be inappropriate, as these may attract both respondents who are located at this place on the logit scale, but also respondents from other ability groups who do not know or do not care (Kulas, Stachowski, & Haynes, 2008).

The appropriateness of a response scale may be assessed by inspecting the probability curves of items' response categories. The thresholds, introduced in Section 4.1, will be located at the intersections between the probability curves of two adjacent response categories. Figure 10 displays the category probability curves of Item 2. These seem to be appropriate, as an increase in ability ("Person Location") successively makes a higher response category more probable.

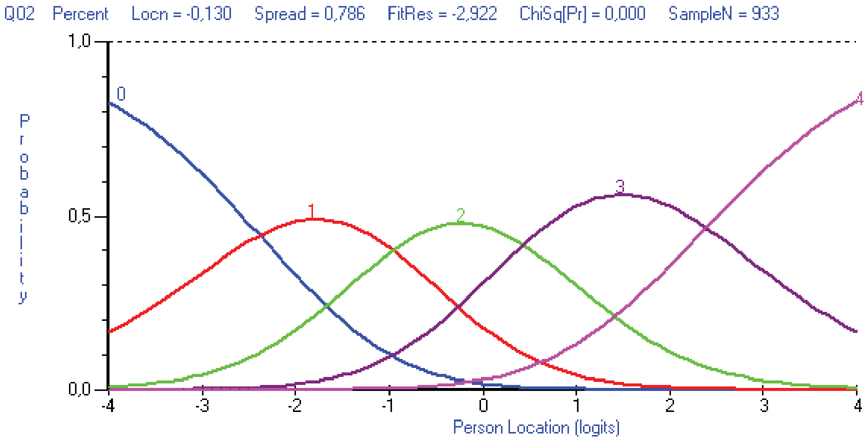


Figure 10. Category probability curves of Item 2. The four thresholds are ordered and the response categories seem to be appropriate. Illustration from Rasch software RUMM 2030.

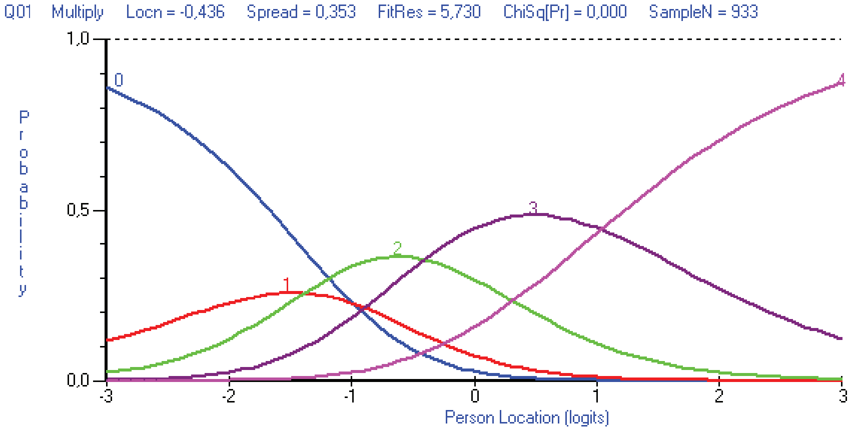


Figure 11. Category probability curves of Item 1. Reversed thresholds for response category 1. Illustration from Rasch software RUMM 2030.

Figure 11 displays the category probability curves of Item 1, which illustrates a case of “reversed” or “disordered” thresholds: The threshold between 0 and 1 is located at -1.0 logit, while the threshold between 1 and 2 is located at -1.5. This result is also detectable in Figure 3, where “01.2” is located before “01.1”, reading from left to right. As a consequence of these disordered thresholds, response category 1 is not the most likely response at any point on the logit scale. Persons increasing in ability tend to move from category 0 directly to category 2 on this item.

Two possible reasons for reversed thresholds were mentioned above, namely, redundant response categories and inappropriate mid-categories. More reasons exist, including instances of multidimensionality. Item 1, displayed in Figure 11, concerns the multiplication table, and respondents seem to move directly from knowing how to apply the table “not at all” to “somehow” as they increase in ability. “Not in particular” is perhaps not a plausible response to a question concerning how well one knows the multiplication table. This response category might be redundant.

4.8 Implications of a Rasch analysis

The properties presented in Section 4.1 throughout Section 4.7 produce statistics which can be applied to identify ways in which data may deviate from expected patterns derived from the Rasch model. It is important to note that data not fitting the Rasch model does not imply that a test, an item or a respondent should be discarded. Misfitting data rather suggest that further investigations are necessary, whether it concerns the constructs, the items, the response scales, the scoring, the respondents, or the test administration. The outputs from Rasch analyses indicate where to begin such investigations, which may include additional Rasch statistics, and statistics from classical test theory (exemplified below).

Moreover, theoretical considerations must always follow instrument development from beginning to end. Such considerations are integrated parts of the process of item development, data collection and throughout the analysis process (Stenner, Fisher, Stone, & Burdick, 2013). It is decisive that the concepts we intend to measure are theoretically founded. Showing that data fit the Rasch model is indeed not sufficient to establish validity, which Messick describes as “an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment” (Messick, 1989, page 13). Indeed, the implications of a Rasch analysis are not straight-forward. Changes may be necessary even though data fit the Rasch model, and misfitting data do not automatically imply that the instrument must be rejected.

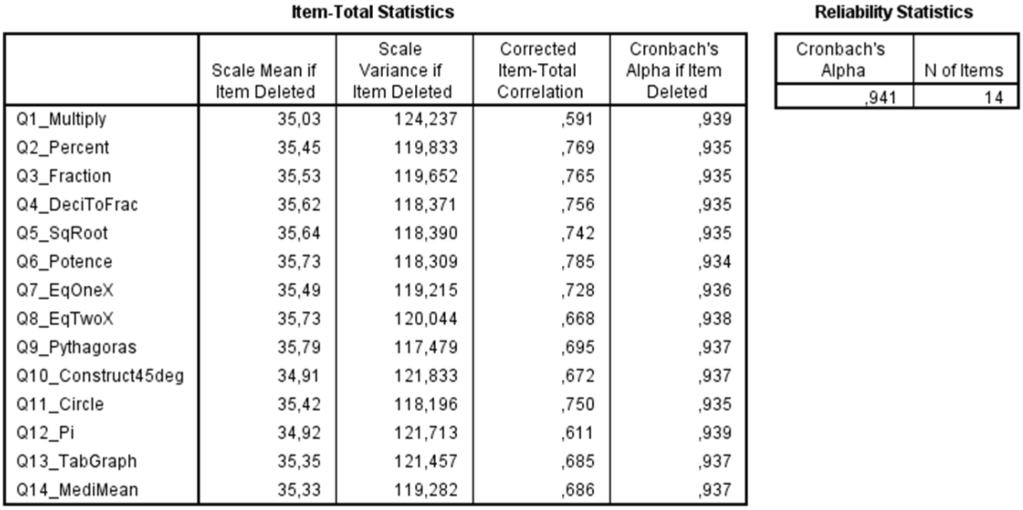
Still, with regard to the NIFU instrument, the analyses reveal some challenges. Suggested by the response pattern of the student presented in Section 4.3, it might be wise to look for response patterns identifying students who did not respond seriously to the survey. A more important concern regards the test difficulty, as a substantial amount of the students responded that they knew most mathematics tasks “very well”. The NIFU survey is thus not tailored to differentiate between the highly skilled students in 9th grade. The immediate solution is to include questions about more challenging mathematics tasks when administered to this group of 9th graders.

The results also suggest further investigation of Item 1 (under-discriminating), Item 2 (over-discriminating) and Item 10 (gender DIF). The latter item was the only item concerning “practical geometry”, a fact related to the dimensionality discussion mentioned in Section 3.2 and fueled by the result in Section 4.6: Is it more appropriate to develop instruments that measure specific mathematics competencies separately? The data were not indisputably unidimensional, and Figure 9 indicated that many persons score significantly different on tasks related to equations and tasks related to fractions, decimal numbers and percentages. One appropriate solution might be to provide different scores to different mathematics competencies. On the other hand, it might still be reasonable to summarize the scores into one overall score – it depends on how the results are utilized, theoretical considerations and supplementing statistics. Either way, the analyses have identified issues in need for a second consideration.

5. What about classical test theory?

Many of the challenges treated in Rasch theory may also be treated using classical test theory (see e.g. Raykov & Marcoulides, 2011). The latter revolves around the concepts of measurement error, true scores and error scores. Many models are derived from the fundamental equation X = T + E, stating that a test score X is the sum of the true score T and the error score E. Compared to Rasch analysis, one can say that the underlying principles are different, but the results often turn out quite similar.

The data from the NIFU survey might be explored with techniques from classical test theory. For instance, while Rasch analysis provided a PSI of 0.92, we can calculate Cronbach's alpha to investigate the reliability. Moreover, "Cronbach's alpha if item deleted" indicates whether or not the different items contribute positively to the measure. The results of these analyses are displayed in Figure 12.



Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
Q1_Multiply	35,03	124,237	,591	,939
Q2_Percent	35,45	119,833	,769	,935
Q3_Fraction	35,53	119,652	,765	,935
Q4_DeciToFrac	35,62	118,371	,756	,935
Q5_SqRoot	35,64	118,390	,742	,935
Q6_Potence	35,73	118,309	,785	,934
Q7_EqOneX	35,49	119,215	,728	,936
Q8_EqTwoX	35,73	120,044	,668	,938
Q9_Pythagoras	35,79	117,479	,695	,937
Q10_Construct45deg	34,91	121,833	,672	,937
Q11_Circle	35,42	118,196	,750	,935
Q12_Pi	34,92	121,713	,611	,939
Q13_TabGraph	35,35	121,457	,685	,937
Q14_MediMean	35,33	119,282	,686	,937

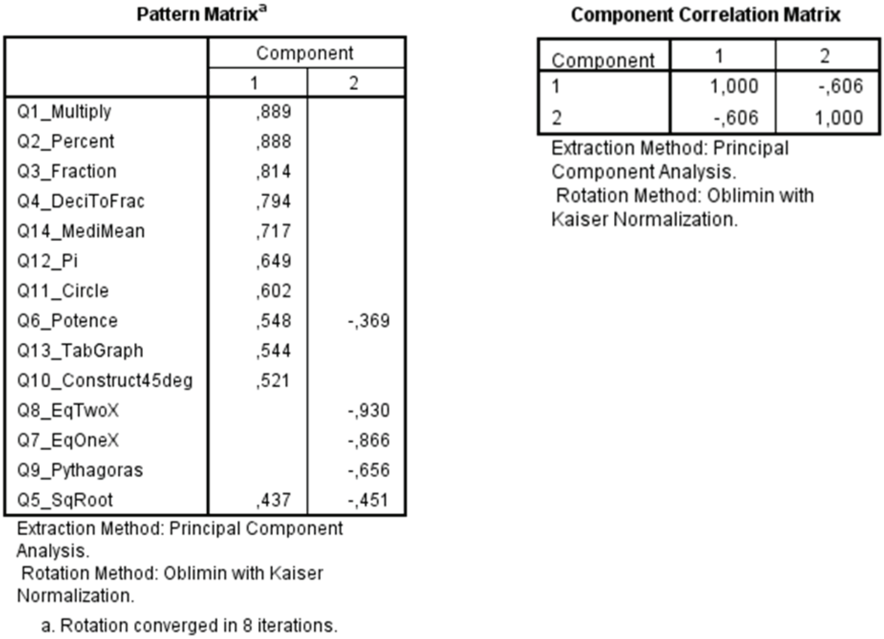
Reliability Statistics

Cronbach's Alpha	N of Items
,941	14

Figure 12. Cronbach's alpha for the full set of items and alpha if item deleted. Illustration from IBM SPSS Statistics 20.

The Rasch analysis resulted in a PSI of 0.92. Cronbach's alpha of 0.94 suggests that the reliability of the collection of mathematics questions in NIFU's survey is good. As the alpha remained below 0.94 for each deletion of an item, no item seems to corrupt the measure. Furthermore, the dimensionality issue was investigated in Rasch analysis drawing on principal component analysis of person-item residuals. In classical test theory, this could also be investigated using a principal factor analysis. Figure 13 displays the results of such an analysis on the data from NIFU's survey using SPSS. As for the results in the foregoing Rasch analysis (Section 4.6), items 2, 3 and 4 load heavily on the first component, while items 7, 8 and 9 load most heavily on the second component extracted in the analysis.

Notably, even though classical test theory provides many results equivalent to those in Rasch analysis, there are good reasons to apply Rasch analysis in instrument development and validation. It provides a variety of important properties like the ones presented in this article, and it does so drawing on fundamental principles of measurement. The software tailored to investigate psychometric properties, concerning issues like differential item functioning and dimensionality, enables researchers to do complex and powerful analyses in efficient and accessible ways. Moreover, some would argue



Pattern Matrix[a]

	Component	
	1	2
Q1_Multiply	,889	
Q2_Percent	,888	
Q3_Fraction	,814	
Q4_DeciToFrac	,794	
Q14_MediMean	,717	
Q12_Pi	,649	
Q11_Circle	,602	
Q6_Potence	,548	-,369
Q13_TabGraph	,544	
Q10_Construct45deg	,521	
Q8_EqTwoX		-,930
Q7_EqOneX		-,866
Q9_Pythagoras		-,656
Q5_SqRoot	,437	-,451

Extraction Method: Principal Component Analysis.
Rotation Method: Oblimin with Kaiser Normalization.

a. Rotation converged in 8 iterations.

Component Correlation Matrix

Component	1	2
1	1,000	-,606
2	-,606	1,000

Extraction Method: Principal Component Analysis.
Rotation Method: Oblimin with Kaiser Normalization.

Figure 13. Factor analysis using principal components and oblimin rotation with Kaiser normalization: Pattern Matrix and Component Correlation Matrix. Illustration from IBM SPSS Statistics 20.

that Rasch theory offers several properties that classical test theory cannot, for instance related to the benefits of locating persons and items on the same interval scale, the model's robustness against missing data, and the calculation of individual person fit statistics (Wright, 1992).

6. Conclusion

The Rasch model is a valuable tool to enhance measurement in science and mathematics education. Rasch analysis supports the development and validation of invariant measures as it provides empirical evidence and insight to important psychometric properties of tests and questionnaires. Seven core properties were introduced and exemplified in this article: Is the test well-targeted? Does the test separate well between the persons? How well do the persons' scores fit the model, and how well do the items fit the test? Do the items function differently for different groups of persons? Does the test produce unidimensional data? Are the response categories appropriate? The answers to these questions will support science and mathematics education researchers as they develop and evaluate instruments with respect to fundamental principles of measurement.

References

Andrich, D. (1978). Application of a psychometric rating model to ordered categories which are scored with successive integers. *Applied Psychological Measurement, 2*(4), 581-594. doi: 10.1177/014662167800200413.

Andrich, D. (1988). *Rasch models for measurement*. Thousand Oaks, California: Sage.

Andrich, D., Lyne, A., Sheridan, B., & Luo, G. (2011). RUMM 2030. Retrived December 15, 2011, from RUMM Laboratory, Perth: http://www.rummlab.com.au/.

Andrich, D., & Marias, I. (Eds.). (2009). *Introduction to Rasch measurement of modern test theory*. Course materiell EDUC8638, semester 2, 2009: The University of Western Australia.

Angell, C., Lie, S., & Rohatgi, A. (2011). TIMSS Advanced 2008: Fall i fysikk-kompetanse i Norge og Sverige [TIMSS Advanced 2008: Fall in physics competence in Norway and Sweden]. *Nordic Studies in Science Education, 7*(1), 17-31.

Bennett, J., & Hogarth, S. (2009). Would you want to talk to a scientist at a party? High school students' attitudes to school science and to science. *International Journal of Science Education, 31*(14), 1975-1998. doi: 10.1080/09500690802425581.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model* (2 ed.). New Jersey, London: Lawrence Erlbaum Associates.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281-302. doi: 10.1037/h0040957.

Ekborg, M., Ottander, C., Silfver, E., & Simon, S. (2012). Teachers' experience of working with socio-scientific issues: A large scale and in depth study. *Research in Science Education, 43*(2), 599-617. doi: 10.1007/s11165-011-9279-5.

Gardner, P. L. (1996). The dimensionality of attitude scales: A widely misunderstood idea. *International Journal of Science Education, 18*(8), 913-919. doi: 10.1080/0950069960180804.

Grimbeek, P., & Nisbet, S. (2006). Surveying primary teachers about compulsory numeracy testing: Combining factor analysis with Rasch analysis. *Mathematics Education Research Journal, 18*(2), 27-39. doi: 10.1007/BF03217434.

Hagquist, C., & Andrich, D. (2004). Is the sense of coherence-instrument applicable on adolescents? A latent trait analysis using Rasch-modelling. *Personality and Individual Differences, 36*(4), 955-968. doi: 10.1016/s0191-8869(03)00164-8.

Kennedy, C., Wilson, M. R., Draney, K., Tutunciyan, S., & Vorp, R. (2011). ConstructMap Downloads. Retrieved December 15, 2011, from Bear Center, Berkeley: http://bearcenter.berkeley.edu/ConstructMap/download.php.

Kulas, J., Stachowski, A., & Haynes, B. (2008). Middle response functioning in Likert-responses to personality items. *Journal of Business and Psychology, 22*(3), 251-259. doi: 10.1007/s10869-008-9064-2.

Linacre, J. M. (2012). Winsteps ministep. Rasch-model computer programs. Retrieved April 16, 2012, from: http://www.winsteps.com/winman/.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (pp. 13-103). London: Collier Macmillian.

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

Neumann, I., Neumann, K., & Nehm, R. (2010). Evaluating instrument quality in science education: Rasch-based analyses of a Nature of Science test. *International Journal of Science Education, 33*(10), 1373-1405. doi: 10.1080/09500693.2010.511297.

Oon, P.-T., & Subramaniam, R. (2012). Factors influencing Singapore students' choice of physics as a tertiary field of study: A Rasch analysis. *International Journal of Science Education, 35*(1), 86-118. doi: 10.1080/09500693.2012.718098.

Oskarsson, M., & Karlsson, K. G. (2011). Health care or atom bombs? Interest profiles connected to a science career in Sweden. *Nordic Studies in Science Education, 7*(2), 190-201.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danmarks Paedagogiske Institut.

Raykov, T., & Marcoulides, G. A. (2011). *Introduction to psychometric theory*. New York: Routledge.

Ryan, J. P. (1983). Introduction to latent trait analysis item response theory. In W. E. Hathaway (Ed.), *Testing in the schools. New directions for testin and measurement*. San Francisco: Jossey-Bass.

Schulz, W., & Fraillon, J. (2011). The analysis of measurement equivalence in international studies using the Rasch model. *Educational Research and Evaluation, 17*(6), 447-464. doi: 10.1080/13803611.2011.630559.

Sjaastad, J. (2012): Measuring the ways significant persons influence attitudes towards science and mathematics. *International Journal of Science Education*, 35(2), pp. 192-212. doi: 10.1080/09500693.2012.672775.

Sjøberg, S., & Schreiner, C. (2010). The ROSE project - An overview and key findings. Retrieved April 16, 2012, from: http://www.roseproject.no/network/countries/norway/eng/nor-Sjoberg-Schreiner-overview-2010.pdf.

Stenner, A. J., Fisher, W. P., Stone, M. H., & Burdick, D. S. (2013). Causal Rasch models. *Frontiers in Psychology, 4*(536).

Stubbe, T. C. (2011). How do different versions of a test instrument function in a single language? A DIF analysis of the PIRLS 2006 German assessments. *Educational Research and Evaluation, 17*(6), 465-481. doi: 10.1080/13803611.2011.630560.

Styles, I., & Andrich, D. (1993). Linking the standard and advanced forms of the Raven's progressive matrices in both the pencil-and-paper and computer-adaptive-tesing formats. *Educational and Psychological Measurement, 53*(4), 905-925. doi: 10.1177/0013164493053004004.

Tennant, A., & Conaghan, P. G. (2007). The Rasch measurement model in rheumatology: What is it and why use it? When should it be applied, and what should one look for in a Rasch paper? *Arthritis Care & Research, 57*(8), 1358-1362. doi: 10.1002/art.23108.

Tennant, A., & Pallant, J. F. (2006). Unidimensionality matters! (A tale of two Smiths?). *Rasch Measurement Transactions, 20*(1), 1048-1051.

Turmo, A., & Elstad, E. (2009). What factors make science test items especially difficult for students from minority groups? *Nordic Studies in Science Education, 5*(2), 158-170.

Wendt, H., Bos, W., & Goy, M. (2011). On applications of Rasch models in international comparative large-scale assessments: A historical review. *Educational Research and Evaluation, 17*(6), 419-446. doi: 10.1080/13803611.2011.634582.

Wright, B. (1992). Raw scores are not linear measures: Rasch vs. classical test theory. *Rasch Measurement Transactions, 6*(1). Retrieved July 16, 2013, from Rasch.org: http://www.rasch.org/rmt/rmt61n.htm.

Wright, B., & Masters, G. N. (1982). *Rating scale analysis*. Chicago: Mesa Press.

Wright, B., & Stone, M. H. (1979). *Best test design*. Chicago: Mesa Press.

Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. (2007). ACER ConQuest Version 2.0 Manual Retrieved 15.12.2011, from https://shop.acer.edu.au/acer-shop/group/CON2.

Zain, A. N. M., Samsudin, M. A., Rohandi, & Jusoh, A. (2010). Using the Rasch model to measure students' attitudes toward science in 'low performing' secondary schools in Malaysia. *International Education Studies, 3*(2), 56-63.

Appendix: Excerpt from the NIFU survey

Main question: "*How well do you master the following tasks in mathematics?*" All items were responded to electronically, using a 5-point Likert scale where the boxes were labeled "not at all" (scored 0), "not in particular" (scored 1), "somehow" (scored 2), "well" (scored 3), and "very well" (scored 5).

1. The multiplication table
2. Calculating percentages
3. Calculating using fractions
4. Calculating decimal numbers into fractions (e.g. 0.5 into ½)
5. Using square roots in calculations
6. Do potency calculations
7. Solve equations with one variable (x)
8. Solve equations with two variables (x and y)
9. Use Pythagoras' sentence in calculating the sides and the area of a triangle
10. Construct a 45 degree angle using math compass and ruler
11. Calculate the circumference and area of a circle
12. Know what the number pi (π) means
13. Understand tables and graphs
14. Explain the difference between median and mean