Curt Hagquist

Karlstad University, Karlstad

Evaluating composite health measures using Rasch modelling: an illustrative example

Summary

Objectives: The purpose of the present article is to elucidate the opportunities provided by Rasch modelling in epidemiology and public health research in order to evaluate composite measures of health.

Methods: The article gives a review of Rasch modelling in conjunction with illustrative examples based on adolescent survey data.

Results: The article demonstrates how the Rasch-model enables examinations of the way items work across different samples/subgroups, e.g., detection of possible differential item functioning.

Conclusions: It is concluded that Rasch modelling may serve as a useful tool in the evaluation and the development of composite health measures intended to be used in epidemiology and public health research.

Key-Words: Rasch models – Health – Measurement.

In public health research, composite quantitative measures play a crucial role, not least when dealing with *perceived* health. The theoretical complexity of some health concepts often makes the sole use of one single indicator insufficient and inappropriate. Hence, in order to capture complex concepts the use of latent measures based upon several indicator variables seems to be straightforward and unquestionable.

In focusing composite measures the dimensionality of the (health) construct turns out to be a key issue to consider. This article will only deal with unidimensional measurement, i.e., health constructs intended to be measured on a single latent variable. In order to achieve uni-dimensional measurement, relevant operating characteristics of the items have to be invariant across individuals, e.g., across different subgroups along the latent trait as well as across different sample groups (e.g., gender). Given that those requirements are met, the health measure will reflect just differences in *degree* of the same *kind.*[1] Ignoring the issue of invariance may cause misinterpretations of the outcomes based on composite measures.

The idea of invariance in the way the instrument worked was recognised by Thurstone[2] as early as in the 1920s. However, Thurstone's method proved cumbersome. Both Likert and Guttman made major contributions to social measurement, but Likert's work was a-theoretical, while Guttman's was deterministic.[3] Independent work by the Danish mathematician Georg Rasch in the 1950s and 1960s, which was theoretically very rigorous and which was set in a statistical framework, addressed the issues of uni-dimensionality and invariance in practice.[4] Rasch stated that "Two features seem indispensable in scientific statements: They deal with comparisons and they must be *objective.*" and introduced the concept of "specific objectivity".[5]

Although Rasch modelling has been frequently used in educational research and is well recognised in connection to social measurement,[6] presentations of Rasch modelling have been slow in being incorporated into standard methodological textbooks. However, the potential of Rasch modelling is now being increasingly recognised within many research fields.

The *purpose* of this article is to elucidate the opportunities provided by Rasch modelling in epidemiology and public health research in order to evaluate composite measures of health.

Soz.- Präventivmed. 46 (2001) 369–378
0303-8408/01/060369-10 $ 1.50 + 0.20/0

Methods

In order to carry out its purpose the article begins with reviews of the basic dichotomous Rasch model and the extended Rasch model for ordered polytomous data respectively. In the next part some illustrative examples of Rasch analyses are provided, in order to demonstrate a typical application of Rasch modelling in public health research. These tentative analyses are based on cross-sectional survey data, collected in February 1988 and April 1998 among students in year nine in the county of Värmland in Sweden. The study was carried out by the County Council of Värmland. The data collection was performed by a questionnaire, which was handed out in the classrooms by school personnel.

For the purpose of this article two minor subsamples from 1988 and 1998 were merged, ending up with a set of data consisting of 535 persons (280 boys and 255 girls). Eight items intended to compound a latent measure of well-being and perceived health are used, that is:

During this school year, have you…
… felt that you have had difficulty in concentrating?
… felt that you have had difficulty in sleeping?
… suffered from headaches?
… suffered from stomach aches?
… felt tense?
… had little appetite?
… felt low?
… felt giddy?

The response categories for all of these questions are "never", "seldom", "sometimes", "often" and "always". Only complete data are used in the analyses, although it would have been possible to use incomplete data.

Two different sample characteristics (person factors) are used in the analyses: gender and year of investigation.

The data analyses are performed using the item analysis program RUMM 2010.[7] In order to estimate the model parameters the programme makes use of a pairwise procedure based on conditional maximum likelihood.[8,9]

The Rasch model: a review

The Rasch model belongs to the class of models emerging from latent trait theory (LTT) or item response theory (IRT). A distinctive feature of the Rasch model is its derivation from theory, i.e., it is constructed a priori to the data.[1] The Rasch model is built upon measurement *requirements*, not on *assumptions* about the data.[10] This implies that the approach taken in Rasch analyses is to compare the data with the model which is considered fixed, since it reflects the required properties of the data. If the data do not fit the model properly, instead of including new parameters the data therefore should be re-examined. Hence, the Rasch model has been considered to be qualitatively different from other response models, giving rise to propositions of paradigm shift.[3]

The ways of viewing the data structure in the Rasch model resemble the pattern analyses proposed by Guttman.[1] The deterministic Guttman pattern helps clarify the essential features of the data, and was used for the same effort by Rasch.[4] The Rasch model as well as the Guttman structure is built upon uni-dimensional scales, i.e., they are intended to measure a single concept represented on a linear continuum. In both models the response pattern is cumulative, e.g., a person scoring high on a severe item is expected to score high on a less severe. Unlike the Guttman structure, in which the items responses are *determined* by the person scores, the persons' locations on the latent trait give rise to the *probabilities* of the item scores.[8] Hence, the Rasch model seems to be more realistic and consistent with human behaviours than the Guttman patterns, since straightforward response patterns without any deviations are very unlikely to occur in real life.

The most unique feature of the Rasch model is that it enables the person and item parameters to be estimated independently of each other, given that the data conform to the model.[4,11] In order to achieve those independent estimates, a sufficient statistic is required that allows the person parameters to be left out when the item parameters is estimated and vice versa. Given that the data fit the Rasch model the sums of the raw scores across items (= person scores) are sufficient statistics for the person parameters and the sums of the raw scores across persons (= item scores) are sufficient statistics for the item parameters.[4] Hence, the total raw scores comprise the bases for the computation of new scores on an interval scale which is common for the person and item parameters.[1] The unit of measurement of these new scale values is log odds ("logits").

So far in this article the Rasch model has been mainly dealt with as being just one single model. In fact the Rasch model is a family of different models.[12] Two main types of models can be ascertained, a dichotomous model and a model for ordered polytomous data. The former is the basic simple logistic model (SLM) while the latter may be viewed as an extension of that model.

Although the illustrations in this article will primarily be focusing on ordered polytomous data, for reasons of introduction it will proceed with the dichotomous model before turning to the extended model.

The dichotomous model

The dichotomous Rasch model contains just two kinds of parameters: the person parameter beta (β) and the item parameter delta (δ). Since the latter one is the only parameter required to "model" the items, the Rasch model is sometimes called "one parameter model", distinguishing it from the "two parameters' model",[13,14] which also contains a discrimination parameter.

The simple logistic model takes the following form:

$$\Pr\{x_{vi}=1\}=\frac{e^{\beta_v-\delta_i}}{1+e^{\beta_v-\delta_i}}.$$

This function can also be represented graphically using item characteristic curves (ICC), one for each item. The curve takes a logistic functional form, i.e., it is s-shaped. The ICC is the expected value curve, i.e., it reflects the item scores that for each person is predicted by the model. In the dichotomous case the ICC is also the probability of a positive response.

The relations between items and persons (which give rise to the response structure) are crucial in the Rasch model: the probability of a specific response becomes a function of the relation between the person parameter estimate and the item parameter estimate, i.e., beta (β) – delta (δ).[15] Positive values from the subtraction imply probabilities above 0.5 for a specific response; the bigger difference the higher probability. Negative values of the subtraction will imply probabilities below 0.5; the bigger difference the lower probability.

The absence of a discrimination parameter in the Rasch model implies that constant discrimination across items is built in as a property of the model. This is justifying the use of raw scores as a sufficient statistic, given that the data conform to the model. Conversely, violations of the requirement of constant item discriminations imply that no sufficient statistic is at hand, which in turn means that invariant comparisons cannot be made.

The frequently used term *discrimination* refers to how the item scores are differentiated across the common scale, i.e. more precisely the rates of change of the expected score relative to the latent measure. The requirement of equal discrimination also means that any increase on the x-axis (i.e., the person location) will imply the same increase on the y-axis (expected value) no matter which item is focused (see Fig. 1). Figure 1 contains an example of items with equal discrimination. Figure 1 shows that the slopes of the four item characteristic curves are parallel, which is indicating equal discrimination. Examined graphically over-discrimination means that the observed scores form a steeper line than the theoretical item characteristic curve generated by the Rasch model. Conversely under-discrimination means that observed scores form a flatter line compared to the theoretical curve. Failure to meet the requirement concerning equal discrimination means a bad fit. In Figure 3 and Figure 4 (see below) examples of over- and under-discriminating respectively are shown.

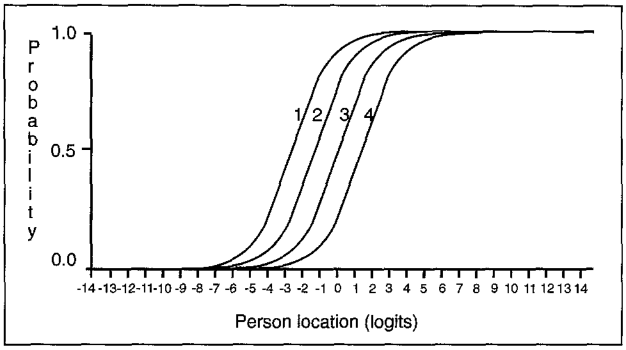

Figure 1 Item characteristic curves for four dichotomous items with equal discrimination

In order to stress the requirements upon which the Rasch model is relying, the phrase "given that the data fit/conform to the model" is used repeatedly in this article. Examining the fit of the model, Rasch himself tended to favour graphical displays instead of formal test statistics[16]. Although graphical representations are most valuable tools in the item analysis process, formal test statistics seem indispensable in order to make statistical statements. The test statistics applicable to the Rasch model emanate from general types of test statistics, e.g., Pearsons chi-squared test, the likelihood ratio tests and the Wald statistics[17]. Different statistics are sensitive to different kinds of violations of the requirements for the Rasch model[18], which means that multiple tests should be carried out in order to evaluate the fit between the data and the model. Furthermore, "Rasch tests" have been questioned with respect to their ability to detect uni-dimensionality[18,19].

The model for ordered polytomous responses

As mentioned above the basic Rasch model for dichotomous responses can be extended to a model for additional categories[11]. As many items (e.g., in health research) contain more than two categories, the model for ordered polytomous data represents an important development of the simple logistic model. The extension of the dichotomous model is straightforward, e.g., the fundamental principles concerning separation of items' and persons' parameters estimates still hold as well as the estimations procedures and the principle of sufficient statistic.

The extended model for ordered polytomous data that Rasch evolved was further clarified by Andersen[20] and Andrich[21]. It takes the following general form:

$$\Pr\{x_{v\mathrm{i}} = x\} = \frac{e^{-\tau_{1i}-\tau_{2i}\ldots-\tau_{xi}+x(\beta_v-\delta_i)}}{\sum\limits_{x'=0}^{m_i} e^{-\tau_{1i}-\tau_{2i}\ldots-\tau_{x'i}+x'(\beta_v-\delta_i)}}$$

Since there are more than two probabilities at each item that have to be calculated additional parameters of a new kind has to be added to the model[21,22]. The number of possible extra parameters reflects the number of thresholds, which in turn are strictly related to the number of categories on the items. The thresholds are the points related to the common logit scale where the *conditional* probability of two adjacent categories (given that the response falls into one of the two categories) occurring is equal (i.e., 0.5). Hence within items with three response categories there are two thresholds, within items with four categories there are three thresholds, and so on. In the extended model each threshold is represented by a parameter τ. The x-value in the numerator is just the score of the item. Since the denominator is the sum of all the single numerators, every single parameter τ will affect the response probability of every x.

The thresholds are located around the item scale values, i.e. the mean value of the estimates of these (centralised) thresholds is zero. Hence, the location of each threshold on the common latent scale is the sum of the estimate of the item parameter and the estimate of the parameter τ (= delta + tau). This means that the thresholds qualify the category values of the items, which enables examinations of the intensity at a category value on one item compared to a numerically equal category value on another item.

Graphically, the thresholds divide the latent scale into different regions that correspond to the categories of the items. The thresholds indicate the form of the ICC, i.e., how steep it is. The closer distances between the thresholds the steeper slope, that is, the stronger discrimination. In contrast to the dichotomous case, the probability value of x is not the same as the expected value of x. In the polytomous case, the response probabilities therefore are displayed separately by category characteristic curves (see Fig. 7 below). On those graphs the thresholds are located at the intersection points of the probability curves at adjacent thresholds.

Similar to the dichotomous model, there is a requirement of equal discrimination also in the model for ordered categories, although that requirement does not apply to the items as a whole but to all of the thresholds of the items. This is reflected by the threshold probability curves, which express a dichotomous response pattern at each threshold.

In order to further elucidate the close link between the basic dichotomous model and the extended model for polytomous ordered data, the idea of thresholds may also be applied into the dichotomous model. In fact, a dichotomous model with two ordered categories can be viewed as a special case of the polytomous model. However, in the dichotomous case there is just one threshold, i.e., where the probability of the two response alternatives is equal. Therefore the threshold is identical to the item scale value which means that no extra parameters are necessary in order to estimate the thresholds.

The model for ordered polytomous data described above allows the thresholds to vary across different items. That feature distinguishes this unrestricted model (sometimes called the partial credit model[23]) from a slightly different variant of the model for polytomous ordered data, the (restricted) rating model[21,22]. In that model the thresholds are set/constrained to be equal across all the items. Hence, the rating model is a more parsimonious variant than the unrestricted model and it contains a less number of parameters, although it is exactly the same *kind* of parameters included in both the models. Since the distances between the thresholds across items become equal in the rating model, the item characteristic curves are parallel, which is similar to the dichotomous model but in contrast to the unrestricted model for ordered data.

Similar to the dichotomous model, the fit of the extended model has to be examined using in principal the same *types* of test statistics[24]. In addition, the items in the extended model have to be examined with respect to the patterns of their categories in relation to the latent measure – in order to make sure that the scores work in an ordered way as expected, i.e., that no reversal scoring occur[25–27]. Reverse threshold ordering indicates problems in the empirical ordering of the response categories. Incorrect threshold ordering is likely to be an indication of non-constant discrimination at the thresholds due to multi-dimensionality in the data. If an item does not discriminate at a threshold and thereby causes reverse ordering of the thresholds, the discrimination of the item as a whole may become exaggerated. A steeper slope of the ICC curve will reflect this in turn.

Although reverse threshold ordering is a violation of the requirement of the ordering of manifest categories, from a numerical point of view the Rasch model does not require a specific ordering of the thresholds. This means that a sufficient statistic is provided and that the ordinary test statistics may indicate "good" fit although the thresholds are incorrect ordered. However, although the thresholds themselves do not have to be constrained in the polytomous model, the response patterns have to. Similar to the dichotomous model the responses have to conform to a Guttman pattern. Unlike the dichotomous case this pattern does not apply directly to the items themselves, but to the thresholds of the items. In the polytomous case the Guttman pattern implies that a person scoring "medium" on an item would have succeeded in meeting the requirements also for scoring "low". Similarly, if a person fail to meet the requirements for scoring "medium", failure is also implied for scoring "high".

In the following, the Rasch model will be further described in conjunction with the illustrative examples.

Tentative analyses and illustrative examples

In Table 1 the frequency distributions of the eight items are shown. Table 1 shows that the prevalence of different health problems as regards the frequency of their occurrence differs substantially. Looking at the "extreme" variables, about one out of five students has experienced *concentration difficulties* often or always but, for example, only about one out of ten has felt giddy often or always.

General level of analysis

In Table 2 the estimates of the item parameters are reported, revealing that the items are representing different degrees of severity. Negative and positive estimates indicate that the items capture relatively less severe (e.g., concentration difficulties) and relatively more severe (e.g., giddy) health problems respectively. Due to the scale construction the values of all the eight item parameter estimates sum to zero.

In order to test if the items meet the model requirement of invariance different options are available, e.g., the chi-squared statistics and graphical representations using item characteristic curves.[8]

The probability values reported in Table 2 are chi-square statistics based on comparisons between observed means and expected values in five approximately equal sized class intervals of persons along the latent trait. These results show that the estimates of all items are statistically non-significant (p > 0.01), i.e., the fit of the items seems good.

Item label	Response category				
	Never (0)	Seldom (1)	Sometimes (2)	Often (3)	Always (4)
Concentrating difficulties	5	24	50	19	3
Sleeping difficulties	20	30	31	16	3
Headache	21	31	28	18	2
Stomach aches	27	35	28	9	1
Tense	20	41	31	8	1
Little appetite	32	37	21	8	2
Felt low	18	38	33	10	2
Giddy	37	31	23	8	1

Table 1 The proportions of responses in different categories for all items used (percent)

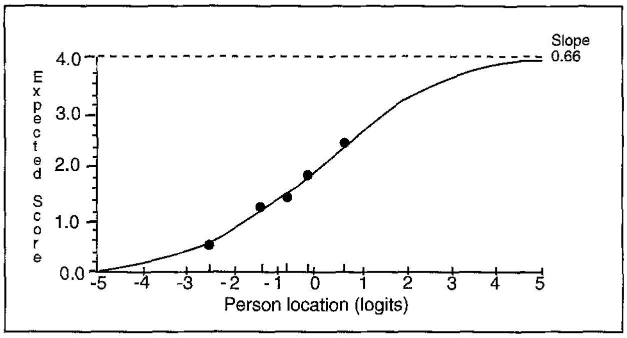

Figure 2 Item characteristic curve for Felt low

Item label	Estimate	Residual	Probability	Thresholds			
				1	2	3	4
Concentrating difficulties	–0.898	0.934	0.535	–3.080	–0.973	1.312	2.741
Sleeping difficulties	–0.214	3.094	0.060	–1.639	–0.645	0.352	1.932
Headache	–0.077	0.118	0.877	–1.842	–0.686	0.129	2.399
Stomach aches	0.216	–0.533	0.717	–1.856	–0.716	0.728	1.844
Tense	0.236	0.000	0.456	–2.434	–0.717	0.809	2.342
Little appetite	0.241	0.172	0.427	–1.629	–0.304	0.487	1.446
Felt low	–0.026	–0.635	0.393	–2.309	–0.627	0.883	2.053
Giddy	0.520	–1.523	0.130	–1.577	–0.826	0.316	2.087

Table 2 Estimates of individual item parameters, item fit and unconstrained centralised thresholds

Similarly to the chi-squared tests the *item characteristic curves* deal with comparisons between observed and expected scores for single items.

Figure 2 contains the item characteristic curve for the item concerned with *felt low*. As would be expected, Figure 2 shows that the higher degree of health problems (measured by the person location) the higher expected scoring on this specific item which measures a single health problem (= *felt low*).

The item characteristic curve in Figure 2 shows the correspondence between the observed scores (calculated as the mean in five approximately equal sized class intervals of persons along the latent trait) and the expected values for the persons in those intervals. The curve demonstrate a good fit, i.e., the dots are located on the line or very close to the line. In addition this apply along the entire x-axis indicating that the item characteristics are invariant across individuals (class intervals) along the latent trait.

An item characteristic curve indicating overdiscrimination is shown in figure 3. This figure shows the item characteristic

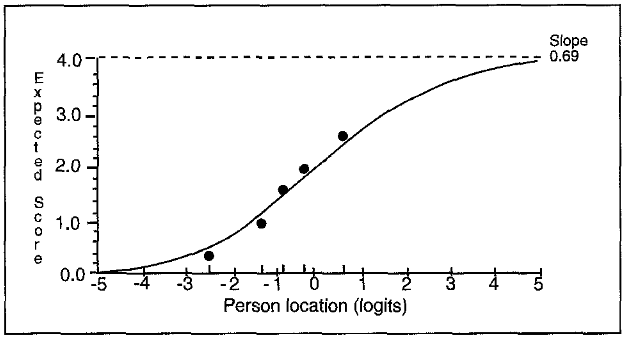

Figure 3 Item characteristic curve for giddy (constrained thresholds)

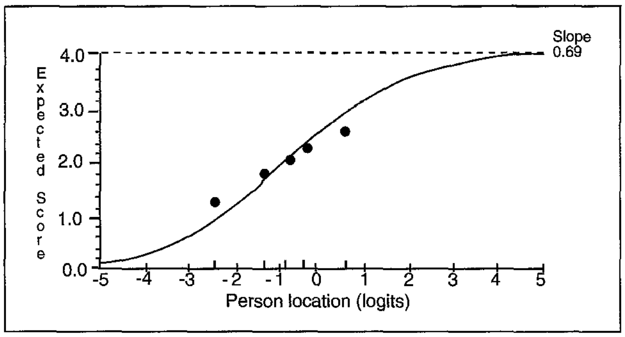

Figure 4 Item characteristic curve for concentration difficulties (constrained thresholds)

curve for *giddy* in the case when the thresholds are constrained to be equal across the items (= the rating model). In relation to the predicted line the observed values form a sharper line, i.e., some over-discrimination occurs. That is, students with no health problems or less severe health problems tend to score too low on this particular item, while students with more severe health problems tend to score too high. In Rasch modelling such over-discrimination is interpreted as a sign of misfit. This view is opposite to the practice based on traditional test theory[28].

An item characteristic curve indicating the opposite pattern, i.e., slightly too poor discrimination is shown in Figure 4. This Figure shows the item characteristic curve for *concentration difficulties* in the case when the thresholds are constrained to be equal across the items. In relation to the predicted curve the observed values form a flatter line, i.e., slightly too poor discrimination occurs. That is, students with no health problems or less severe health problems tend to score too high, while students with more severe health problems tend to score too low.

Item label	**Probability values**					
	Division by gender			**Division by year of investigation**		
	Gender	**Class interval**	**Gender by class interval**	**Year**	**Class interval**	**Year by class interval**
Concentrating difficulties	0.070	0.509	N/Sig	0.285	0.509	0.788
Sleeping difficulties	0.013	0.059	0.892	0.194	0.060	0.646
Headache	0.376	0.836	0.405	0.011	0.832	0.194
Stomach aches	0.001	0.583	0.944	0.926	0.591	0.543
Tense	0.261	0.391	0.968	0.200	0.387	0.415
Little appetite	0.056	0.408	0.667	0.895	0.412	0.667
Felt low	0.000	0.216	0.124	0.818	0.252	0.548
Giddy	0.063	0.083	0.114	0.394	0.086	0.349

Table 3 Detection of differential item functioning using analysis of variance of standardised residuals. Probability values based on F-ratios for each item and divided according to gender and year of investigation respectively

Finer level of analysis

Although the overall correspondence between the health measure and the items looks fine, distortions may be obscured in the data and become detectable at a finer level of analysis. The possible lack of invariance should therefore be examined not just along the latent trait but also across subgroups like gender, using graphical representations as well as formal test statistics.

Most efficient may be to simultaneously analyse item-by-latent trait and item-by-person factor interaction. Analysis of variance (based on the standardised residuals) is an appropriate method for this purpose, since it enables the test of fit to be partitioned into main effects and interaction effects[29].

Table 3 shows the results from analysis of variance for all the items, separately performed for gender and year of investigation respectively. The results show significant ($p < 0.01$) gender main effects for two items and almost significant main effects for two items (one for gender and one for year). With respect to item-by-latent trait (class intervals) no items show significant effects, neither when divided by gender nor year of investigation. Furthermore, no gender-by-class interval or year-by-class interval interaction effects occur.

In Figure 5 item characteristic curves for boys and girls with respect to *felt low* are shown. In contrast to Figure 2 (describing a *felt low* curve for the entire sample), Figure 5 demonstrates poor fit, i.e., the responses are not invariant between the two subgroups. More precisely, the figure indicates gender differences with respect to the student's reports of *felt low*. Given the same overall health problem (= person

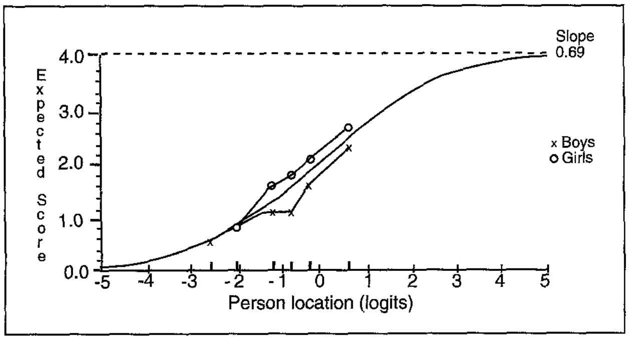

Figure 5 Item characteristic curves for felt low, divided by gender

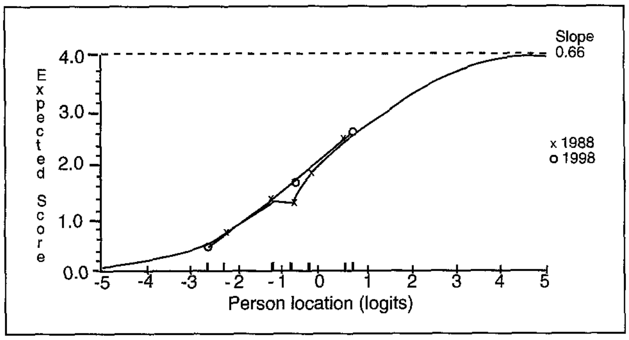

Figure 6 Item characteristic curves for felt low, divided by year of investigation

location), in most intervals the girls score *felt low* to a higher degree than boys do. This will affect the gender differences in person measurement, all others being equal.

Figure 6 shows item characteristic curves for 1988 and 1998 separately with respect to *felt low*. In contrast to Figure 5, Figure 6 illustrates a good fit between observed scores and expected values for both years, indicating invariance and lack of item by year interaction.

As mentioned above, in Rasch modelling based on three or more categories the items do not have just to be evaluated with respect to the correspondence between observed scores and expected values but also with respect to the empirical ordering of the categories. One way of doing this evaluation is to examine the parameter τ, i.e., the values of the thresholds.

In Table 2 the thresholds for all items are shown. The table shows that the thresholds within each of the eight items are successively ordered, which indicates that the items work properly in this sense. A graphical way of examining the ordering of the thresholds is to make use of category probability curves. In Figure 7 such curves are shown for *felt low*.

The shapes and patterns of the curves in figure 7 indicate that the categories work properly, i.e., in an ordered way.

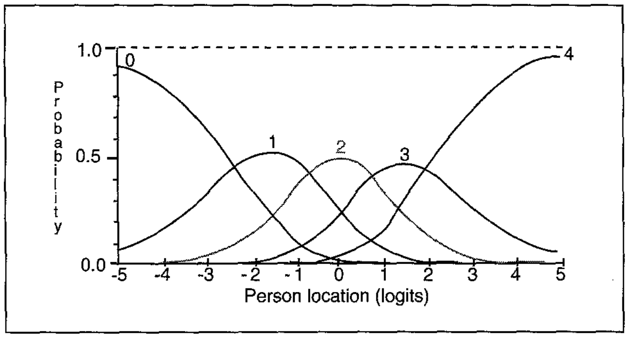

Figure 7 Category probability curve for felt low

Having a low (negative) value on the health scale indicates a high probability of scoring on the lowest value on the items. Conversely, having a high (positive) value on the overall measure, the probability of scoring a high value on the single item is high. Describing this in another way, the thresholds seem to work well, with values ranked successively from low to high.

Summarising the tentative analyses, the Swedish data turn *as a whole* out to fit the unconstrained Rasch model for ordered categories fairly well. Since the finer level of analysis turned up with some item-by-person factor interactions, judgements have to be made as to how to handle those items suffering differential item functioning. The most drastically option might be to discard the "bad" items. However, as a rule of thumb, deletions of items should primarily be guided by theory. Therefore, appropriate methods to take differential item functioning into account, i.e., allowing the items to be kept in the model, should be considered. As a result the precision of measurement is likely to be enhanced[30].

Discussion

The theoretically derived Rasch model is based on measurement requirements of the data, which in turn is built in as properties of the Rasch model. Therefore, the model should be used for examination of data, not for description of data.

The Rasch model is obviously at the same time both simple and sophisticated. The simplicity of the Rasch model lies in the minimum number of parameters used in the model equation. The absence of a discrimination parameter may be considered to be a weakness, making the model too restrictive. Thus, if the view is taken that a model should be used primarily to describe the data then the parsimonious parameterisation will mean that the Rasch model sometimes will be difficult to fit. However, the Rasch model enables detection of bad fit which otherwise would be obscured by the discrimination parameter, i.e., absorbed as a property of the item.[3] Furthermore, from the point of view of understanding the data, failure to fit the model may be very informative. Most important, the Rasch model is the only latent trait model with a sufficient statistic[31], enabling objective and invariant comparisons. Therefore, the following view of Molenaar[32] might be an appropriate guideline for the process of item analyses: "Whenever possible, it is thus recommended to find a set of items that satifies the RM [= Rasch Model, CH], rather than find an IRT [Item Response Theory, CH] model that fits an existing item set."

Using composite measures of health, it is important to examine the possible impact of the respondents' reference frames on their responses. To the degree that the data fit the Rasch model, to that degree the measures of the persons' effects can be freed from the impact from the respondents' reference frames. Hence, the requirement of invariance is decisive in order to achieve fundamental measurement. It is important to make sure that the items work consistently for the individuals regardless of the severity of their health problems. Similarly, it is important that the items work consistently across different sample groups that are to be compared. For example, in gender analyses the invariance requirement is necessary to ensure that males and females view health in a similar way; in trend analyses to ensure that the views of health have not changed over time; in comparisons between countries to ensure that health is viewed in a similar way across different cultures.

From a technical point of view, the Rasch model also offers a solution to a common problem arising when outcomes based on composite measures are analysed. Simply adding scores from different items[33] *may* be misleading, if the items and the respondents' raw scores are not related in a similar way to the construct in question.[34] However, given that the data fit the Rasch model the actual response structure can be ignored since it is accounted for by the model. This also means that neither uncertain assumptions about equal item difficulty and equal distances between the response categories nor weights on the items have to be of concern. Furthermore, since the Rasch model provides scores on an interval scale, the outcomes provided by the Rasch analyses may not only be used for item calibration but also for person measurement based on parametric statistics.

To take optimal advantage of Rasch modelling it should be integrated as a part of the early process of developing scales and instruments, facilitating the creation of theoretically robust constructs. However, even at the later stage of analyses (i.e., after the data collection has taken place) invariance between different subgroups may be explored, interpreted or adjusted for in a constructive way using Rasch modelling. For example, items interacting with person factors like gender may be retained without changing the fundamental specifications of the model[30].

Hence, not least within epidemiology and public health research Rasch modelling may serve as a useful tool for development and assessment of questionnaires.

Acknowledgements

Many thanks to Professor David Andrich for supporting my work with this article and for the most valuable comments on earlier drafts. My thanks also to Dr Barry Sheridan for computer support and to Dr Magnus Stenbeck for valuable comments on an earlier draft.

Zusammenfassung

Auswertung zusammengesetzter Gesundheitsmasse nach dem Rasch-Modell: ein erläuterndes Beispiel

Fragestellung: Zweck des vorliegenden Artikels ist es, die Möglichkeiten des Rasch-Modells in der epidemiologischen und Public-Health-Forschung zwecks Auswertung zusammengesetzter Gesundheitsmasse zu erläutern.

Methoden: Der Artikel bietet einen Überblick über das Rasch-Modell in Verbindung mit erläuternden Beispielen, die auf statistischen Daten von Erwachsenen basieren.

Resultate: Der Artikel zeigt, wie das Rasch-Modell die Möglichkeit bietet, die Wirkung einzelner Fragen bei verschiedenen Auswahl- bzw. Untergruppen zu überprüfen, d. h. die Aufdeckung unterschiedlicher Funktionen bei Fragen zu ermöglichen.

Schlussfolgerungen: Als Schlussfolgerung gilt, dass das Rasch-Modell als geeignetes Instrument bei der Auswertung und Entwicklung zusammengesetzter Gesundheitsmasse dienen kann, die in der epidemiologischen und Public-Health-Forschung verwendet werden sollen.

Résumé

Évaluation de mesures de santé composites selon le modèle de Rasch: un exemple illustratif

Objectifs: L'objectif de cet article est d'éclairer les possibilités offertes par le modèle de Rasch en matière d'épidémiologie et de recherche en santé publique afin d'évaluer des mesures composites de santé.

Méthodes: L'article étudie le modèle de Rasch apartis de données provenant d'études avec les adolescents.

Résultats: Il montre comment le modèle de Rasch permet de comprendre la manière dont les différentes variables fonctionnent dans divers échantillons, c'est à dire de détecter le fonctionnement des écarts éventuels entre les differentes variables.

Conclusions: En conclusion, le modèle de Rasch peut constituer un outil très utile dans l'évaluation et le développement de mesures de santé composites utilisées en épidémiologie ainsi qu'en recherche dans le domaine de la santé publique.

References

1 *Andrich D.* An elaboration of Guttman scaling with Rasch models for measurement. In: Brandon-Tuma N, ed. Sociological methodology. San Fransisco: Jossey-Bass, 1985: 33–80.

2 *Thurstone LL.* Attitudes can be measured. Am J Sociol 1928; *33*: 529–54.

3 *Andrich D.* A scientific revolution in social measurement. Paper presented at the first meeting of the American Educational Research Association's Special Interest Group on Rasch Measurement. New Orleans, April 1988.

4 *Rasch G.* Probabilistic models for some intelligence and attainment tests. (First published 1960 by the Danish Institute for Educational Research). Chicago: MESA Press,1980.

5 *Rasch G.* An informal report on the present state of a theory of objectivity in comparisons. In: Van der Kamp LJT, Vlek CAJ, eds. Psychological measurement theory: proceedings of the NUFFIC international summer session in science at "Het Oude Hof", The Hague, July 14–28, 1966. Leyden: University of Leyden, 1967: 1–19.

6 *Duncan OD.* Notes on social measurement historical and critical. New York: Russel Sage Foundation, 1984.

7 *Andrich D, Sheridan B, Luo G.* RUMM2010: a windows interactive program for analysising data with Rasch Unidimensional Models for Measurement. Perth, RUMM Laboratory, 2000.

8 *Andrich D.* Rasch Models for measurement. Newbury Park: Sage, 1988.

9 *Andrich D, van Schoubroeck L.* The General Health Questionnaire: a psychometric analysis using latent trait theory. Psychol Med 1989; *19*: 469–85.

10 *Andrich D.* Distinctions between assumptions and requirements in measurement in the social sciences. In: Keats JA, Taft R, Heath RA, Lovibond SH, eds. Mathematical and theoretical systems. North Holland: Elsevier Science Publishers BV, 1989: 7–16.

11 *Rasch G.* On general laws and the meaning of measurement in psychology. Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability. Berkely: University of California Press, 1961: 321–33.

12 *Wright BD, Mok M.* Rasch models overview. J Appl Meas 2000; *1*: 83–106.

13 *Ryan JP.* Introduction to Latent Trait Analysis and Item Response Theory. In: Hathaway WE, ed. Testing in the schools: new directions for testing and measurement, 19. San Fransisco: Jossey-Bass, 1983: 49–65

14 *Streiner DL, Norman GR.* Health measurement scales: a practical guide to their development and use. Oxford: Oxford University Press, 1995.

15 *Wright BD, Stone MH.* Best test design. Chicago: Mesa Press, 1979.

16 *Andersen EB.* What Georg Rasch would have thought about this book. In: Fischer GH, Molenaar IW, eds. Rasch models Foundations, recent developments, and applications. New York: Springer, 1995: 383–90.

17 *Glas CAW, Verhelst ND.* Testing the Rasch model. In: Fischer GH, Molenaar IW, eds. Rasch models Foundations, recent developments, and applications. New York: Springer, 1995: 69–95.

18 *Van den Wollenberg AL.* Testing a latent trait model. In: Langeheine R, Rost J, eds. Latent trait and latent class models. New York: Plenum Press, 1988: 31–50.

19 *Hattie J.* Methodology review: assessing unidimensionality of tests and items. Appl Psychol Meas 1985; *9*: 139–64.

20 *Andersen EB.* Sufficient statistics and latent trait models. Psychometrika 1977; *42*: 69–81.

21 *Andrich D.* A rating formulation for ordered response categories. Psychometrika 1978; *43*: 561–73.

22 *Andrich D.* A model for contingency tables having an ordered response classification. Biometrics 1979; *35*: 403–15.

23 *Masters GN.* A Rasch model for partial credit scoring. Psychometrika 1982; *47*: 149–74.

24 *Glas CAW, Verhelst ND.* Tests of fit for polytomous Rasch models. In: Fischer GH, Molenaar IW, eds. Rasch models Foundations, recent developments, and applications. New York: Springer, 1995: 325–52.

25 *Andrich D, de Jong JHAL, Sheridan BE.* Diagnostic opportunities with the Rasch Model for ordered response categories. In: Rost J, Langeheine R, eds. Applications of latent trait and latent class models in the social sciences. Münster: Waxmann, 1997: 59–70.

26 *Andrich D.* A general form of Rasch's extended logistic model for partial credit scoring. Appl Meas Educ 1988: *1*: 363–78.

27 *Andrich D.* Measurement criteria for choosing among models with graded responses. In: von Eye A, Clogg CC, eds. Categorical variables in developmental research. Methods of analysis. San Diego: Academic Press, 1996: 3–35.

28 *Masters GN.* Item discrimination: when more is worse. J Educ Meas 1988; *25*: 15–29.

29 *Glass GV, Stanley JC.* Statistical methods in education and psychology. New Jersey: Prentice-Hall, 1970.

30 *Andrich D, Hagquist C.* Taking account of differential item functioning through principles of equating. Perth: Social Measurement Laboratory, Murdoch University, 2001. (Research report; no 12, April 2001)

31 *Heinen T.* Latent class and discrete latent trait models: similarities and differences. Thousand Oaks: Sage, 1996.

32 *Molenaar IW.* Some background for item response theory and the Rasch model. In: Fischer GH, Molenaar IW, eds. Rasch models Foundations, recent developments, and applications. New York: Springer, 1995: 3–14.

33 *Likert R.* A technique for the measurement of attitudes. Arch Psychol (New York) 1932.

34 *Duncan OD, Stenbeck M.* Are Likert scales unidimensional? Soc Sci Res 1987; *16*: 245–59.

Address for correspondence

Curt Hagquist, PhD
Karlstad University
SE-651 88 Karlstad

e-mail: Curt.Hagquist@kau.se