AUTHORS Grondin, Julie; Blais, Jean-Guy

TITLE A Rasch analysis on collapsing categories in item's response scales of survey questionnaire: Maybe it's not one size fits all

PUBLICATION DATE May 1st, 2010

NOTE Paper presented at the annual meeting of the American Educational Research Association (Denver, CO, April 30- May 4, 2010).

ABSTRACT

When respondents fail to use response scales of survey questionnaires as intended, latent variable modeling of data can produce disordered category thresholds. The objective of this paper is to show the usefulness of the Rasch modeling features to explore different ways of collapsing categories so that they are properly ordered and fit for further analysis. Twenty-four items of a survey questionnaire with a response scale composed of six categories were analyzed. Among the many strategies explored and those suggested as guidelines by researchers, one provided much better results. It appears that suggested guidelines regarding how to collapse categories are just guidelines and should not be applied blindly. As a matter of fact, they should be related to each context.

A Rasch analysis on collapsing categories in item's response scales of survey questionnaire: Maybe it's not one size fits all

Julie Grondin
Université du Québec à Rimouski, Campus de Lévis

Jean-Guy Blais
Université de Montréal

1. Introduction

Since Likert's (1932) introduction of the summative method for the measurement of attitudes, ordered response scales have enjoyed great popularity among social sciences researchers who use them to measure not only attitudes and opinions about various phenomena, but also for many other purposes including the assessment of a person's performance and/or ability (Davies, 2008). While the extensive use of these response scales for assessing participants' attributes and answers to survey questionnaires has contributed to obtain better knowledge on many topics of social relevance, it has also drawn research attention to the effects the scale format can have on the responses given as well as on the associated psychometric properties (Weng, 2004).

Researchers in the field of survey questionnaires are well aware that a careful response scale design is essential to achieve satisfactory scale reliability and appropriate research conclusions. Among the topics related to response scale design, the issue of how the number of response categories affects scale reliability is an intensively examined one. It is also well known that the number of categories can influence answers given in self-report instruments (Kirnan, Edler, & Carpenter, 2007) and have profound effects on both the cognitive response burden and the sensitivity of the scoring design (Hawthorne, Mouthaan, Forbes, & Novaco, 2006). But studies examining this issue have produced conflicting results (Chang, 1994; Weathers, Sharma, & Niedrich, 2005; Weng, 2004).

According to Poulton (1989), there should be about five or fewer response categories in order for a respondent to be able to perform his or her task more or less perfectly. But, since reliability generally seems to increase with the number of categories offered, some researchers think that there should be more than five and up to seven or nine categories (and even up to 11) offered. Indeed, it is a common

belief that more scale points will generally be more effective than fewer points as more refined response categories allow respondents to endorse a category which describes his or her attitude or opinion more accurately. More scale points also have the potential to convey more useful information and allow researchers to better discriminate between respondents' attitudes/opinions (Krosnick & Fabrigar, 1997; Weng, 2004). However, it has also been shown that too many response options may reduce the clarity of meaning. As the number of scale points increases, respondents must discriminate between finer response categories which increase the complexity of the task. Respondents may then fail to distinguish reliably between adjacent categories (Weng, 2004). This may lead to less consistency within and between individuals regarding the meaning respondents give to each response option (Wright & Linacre, 1992). So how many anchor points should be included in a response scale? Further investigation on the subject seems warranted (J. Dawes, 2007).

According to Dawes (2007), most of survey data are not just reported. Rather, they are analyzed with the objective of "explaining" a dependent variable. This usually means that researchers will use some sort of overall score on the dependent variable and then try to find out if other variables might be strongly related to higher or lower scores on that variable. Data are thus analyzed as if they were equal-interval. This may be a quick and easy way of analyzing the data, but it generally disregard the subjective nature of the data by making unwarranted assumptions about their meaning. Based on the a priori arrangement of the response categories, as presented in the questionnaire used, these methods are counterintuitive and mathematically inappropriate to analyze Likert-type scales (Bond & Fox, 2001).

When researchers produce this kind of overall score, they presume a ratio, or at least an interval scale for their data. As a result, the relative value of each response category is treated as being the same, and the unit increase across the rating scale are given equal value. Also, each item is considered in the exact same way so that each one contributes equally to the overall score. However, the "real" locations of the thresholds generally do not corroborate this traditional assumption. Likewise, the items of a survey questionnaire usually do not carry the same relative value in the construct under investigation.

It therefore seems appropriate to look for a model that would allow an analysis with finer details of the item and scale structures. This is exactly what the Rasch Rating Scale model developed by Andrich (1978) does: it provides both an estimate for each item as well as a set of estimates for the thresholds that mark the boundaries between the categories in the scale. As Bond and Fox mentions (2001), the model explicitly recognizes the scale as ordered categories only, where the value of each category is higher than of the previous one, but by an unspecified amount. That is, the data are regarded as ordinal (not interval or ratio) data. Also, the model transforms the counts of the endorsements of these ordered categories into interval scales based on the actual empirical evidence, rather than on some unfounded assumption made beforehand. Consequently, the Rasch model analysis of data from Likert-type items in opinion/attitude questionnaire is intuitively more satisfactory and mathematically more justifiable than the traditional approach of the summative method.

One objective of this paper is thus to show how the Rasch model can contribute to explore different strategies to collapse categories when disordered thresholds occur in response scales used in survey questionnaires. The main focus of this paper is related to the different ways categories can be collapsed in order for the data to fit the model optimally. It should be noted that this article follows the trail of previous work done on Likert-type response scales used with items in survey questionnaires with the help of Rasch models. Some results were presented at IOMW 2004 in Cairns, Australia; others were presented at AERA 2007 in Chicago and AERA 2008 in New York.

2. The number of response categories in rating scales

There are two main issues to consider in the development of parsimonious measurement instruments : the number of items to include in the questionnaire and the number of item response categories that will be provided to respondents (Hawthorne et al., 2006). Generally, parsimony in regard to the number of items is well understood: too many items means longer time taken to answer and some impact on answers' reliability. But when it comes to the second issue, parsimony is harder to reach. First, there should be enough categories offered in the item's response scale for a respondent to endorse a category which accurately describes his or her situation. Also, there are various reasons to believe that more scale points will generally be more effective than fewer. This is because people's perceptions of their attitudes/opinions presumably range along a continuum, going

from extremely positive to extremely negative (Krosnick & Fabrigar, 1997), and the set of options offered should represent this entire continuum.

There are a number of theoretical issues researchers should consider before deciding on the number of scale points to include along that continuum (Krosnick & Fabrigar, 1997). First, rating scales can be structured as either bipolar or unipolar (Schaeffer & Presser, 2003). Bipolar scales are used to reflect two alternatives that are in opposition along the continuum, and separated by a clear conceptual midpoint that makes the transition from one side to the other. Attitudes/opinions can usually be thought of as bipolar constructs and, as a matter of fact, bipolar scales are probably the most common scale type used in questionnaires targeting attitudes/opinions (R. M. Dawes & Smith, 1985). In contrast, unipolar positive scales are used to reflect different levels (frequencies, importance) on a given continuum, with no conceptual midpoint, but with a zero at the beginning of the scale.

A second issue when considering bipolar scale is its midpoint, particularly since it can be given different meaning that will influence the responses provided by participants. A rating scale midpoint can be conceived of as indicating indifference (e.g., neither boring nor interesting) or as ambivalence (e.g., boring in some ways and interesting in others) (Schaeffer & Presser, 2003). According to Klopfer and Madden (1980), the middle category is the response option that typically reflects indecision and the three processes that can determine this choice are ambivalence, neutrality and uncertainty. The definition a researcher decides to give to the midpoint may affect the meaning of the other points on the scale (Schaeffer & Presser, 2003). Also, it has been shown that for some constructs, the label used for the middle category may affect how often it is chosen. As an example, more respondents chose the middle category when it was labelled “ambivalent” than when it was labelled “neutral” when the task was to rate capital punishment (Klopfer & Madden, 1980).

From here, it is possible to take an additional step and try to make a distinction between scales that do propose a midpoint, i.e. scales that have an odd numbers of scale points, from those that do not, i.e. that have an even number of points. In other words are attitudes/opinions best recorded as «agree/neutral/disagree» or as «strongly agree/agree/disagree/strongly disagree»? Respondents that have no attitude/opinion toward an object, that have an ambivalent feeling, or that are uncertain, would

presumably try to place themselves at the middle of the continuum offered by the scale. However, if they are faced with a rating scale with an even number of response options, there is no midpoint that would reflect their situation forcing them to choose between a weakly positive or a weakly negative attitude/opinion. This choice may often be random. Consequently, scales with odd numbers of response categories may be more reliable than scales with even numbers of response alternatives because they simply represent reality in a better way (Cools, Hofmans, & Theuns, 2006). Alwin and Krosnick (1997) tried to verify this hypothesis and they found that two- and four-points scales were more reliable than a three points scale. On the other hand, they found that a five points scale was no more reliable than a four-points one. It is also often hypothesized that when a midpoint option is offered respondents may easily adopt a “satisfycing” strategy, i.e. they may seek for a quick and satisfactory response rather than for an optimal one. If this is the case, the midpoint may be chosen more often and as a result, scales would seemingly show greater reliability (Cools et al., 2006). Again, studies examining the relation between reliability and the number of response categories in a scale have produced conflicting results.

Once the researcher has decided the type of scale (bipolar or unipolar), odd or even number of categories, and the meaning of the midpoint if necessary, the researcher must still decide how many response categories to include (and the semantic describing each response option). As an example, a rating scale using only three options and a semantic consisting of “agree”, “disagree” and “neutral” can be considered. But, such a scale does not allow people to say that they agree slightly to something. A respondent agreeing slightly to something is confronted with a difficult decision : choosing the “agree” category, which may imply stronger positivity than it is the case, or to select the “neutral” category which may imply some kind of indifference, uncertainty or ambivalence, while it is not necessarily concordant with his or her situation (Alwin & Krosnick, 1991; Krosnick & Fabrigar, 1997). Offering respondents relatively few responses options may therefore not provide enough scale differentiation for respondents to express reliably their situation and the choices respondents make may very likely be random (Alwin & Krosnick, 1991). This also raises the question whether reliability is affected or not by such imprecision. Consequently, although with only few response options the meaning of the response categories is quite clear, it seems that using more response options would allow

respondents to express their attitudes/opinions more precisely and comfortably (Krosnick & Fabrigar, 1997).

Generally, rating scales with four to 11 response categories are used (Cools et al., 2006) and, historically, five response categories scales have been the convention for self-report instruments (Hawthorne et al., 2006). According to Weng (2004), a scale with fewer than five response options should, if possible, be discouraged because some research results show that the reliability estimates seemed to fluctuate from one sample to another. Cools, Hofmans and Theuns (2006) also found that a five response options scale was least prone to context effect and that supplementary extreme answers, such as “fully agree” or “fully disagree” did not improve the metric properties of the scale. Finally, considering that respondents may only naturally be able to distinguish between slight and substantial leaning, both positively and negatively, a five-point scale might be optimal (Krosnick & Fabrigar, 1997).

Arguments for more response categories relies on the idea that people may be inclined to think of their situation as being either slight (weak), moderate or substantial (strong) for both positive and negative evaluations (Alwin & Krosnick, 1991; Krosnick & Fabrigar, 1997). This is probably because these are the categories that people often use to describe their attitudes and opinions. According to Alwin and Krosnick (1991), a seven-points response scale does seem preferable to shorter ones and, when they are fully labelled, they should increase the likelihood of inducing a stable participant reaction on the measures making them more reliable than those not so labelled (Alwin & Krosnick, 1991; Weng, 2004). In the studies they reviewed, Krosnick and Fabrigar (1997) also found that the reliability was greater for scales with approximately seven points.

Although seven response categories could be the optimal number of response options on a scale, respondents would probably need a much bigger range of options to cover their entire perceptual range (Borg, 2001). Increasing the number of response categories could thus enable respondents to map their situation to the appropriate category which may reduce random error and raise reliability. But there is a limit to the benefit of adding response categories. Indeed, this limit is related to channel capacity limitation, i.e. the ability to meaningfully discriminate between different choices (Hawthorne et al., 2006). According to Miller (1956), people can reliably discriminate between seven categories, plus

or minus two. Once scales grow much beyond seven points, the meaning of each response category may become too ambiguous for respondents to be able to perform their task. Also, some categories may tend to be underused, especially when as many as nine response options are offered (Cox, 1980). As an example, Hawthorne, Mouthaan, Forbes and Novaco (2006) found that the nine categories response scale they used may have confused their participants, suggesting that fewer categories may work equally well or better. Similarly, Cook, Amtmann and Cella (2006) did not find any advantage in measuring pain using 11 categories. Indeed, individuals may have difficulty discriminating the difference between 8 and 9 on an 11-points scale (Weng, 2004). A respondent may choose 8 on one item and 9 on another occasion for an identical item. This inconsistency would then be due to scale design rather than to the trait being measured. Also, the ambiguity created by too many response options is likely to increase random measurement errors (Alwin & Krosnick, 1991; Weng, 2004).

So it seems that the optimal number of response alternatives would be a scale that is refined enough to be able to transmit most of the information available from respondents, but without being so refined that it simply encourages response error (Cox, 1980). A response scale using from five to nine categories should therefore produce relatively good results.

In most cases, once a response scale has been defined, the same scale is applied to all items of a questionnaire. Several dimensions can be relevant for questions on attitudes/opinions, but researchers are often interested in only one or two of the dimensions that relates to the attitude/opinion under investigation. Moreover, respondents would probably not tolerate being asked about all the dimensions at once (Schaeffer & Presser, 2003). Finally, in order to easily be able to create an overall score that summarizes a respondent's answers, each item of a questionnaire is generally designed to contribute equally to the measurement of the selected dimensions of the attitude/opinion being measured and the same response scale is applied to all items. Indeed, such a summative method would not make much sense if the response scale was different from one item to another. But is this the best model to use to study the respondents' answers? If the traditional approach hardly accommodate for different number of categories in the response scales used for each item, it is of no problem for Rasch models. When constructing a rating scale, a researcher habitually

intends to define a clear ordering of response levels. However, for many reasons people often respond differently from what was intended. A researcher may have given more categories than respondents can distinguish or, respondents may answer using multiple dimensions of the attitude/opinion being measured (Andrich, 1996). As it will be shown in the next section, Rasch models can help one verify if the response scale was used according to the intended ordering. Moreover, if two categories were indistinguishable to respondents, Rasch models allow the researcher to combine these two categories to see if the rating scale works more closely to what was intended and if data better fit the model.

3. The Rasch model for Likert-type rating scales

3.1 The model

Rasch models are so named in the honour of Georg Rasch, a Danish mathematician who developed a model to analyze dichotomous data (Rasch, 1960/1980). Almost twenty years later, David Andrich extended the Rasch family of models by developing a model for rating scale data (1978). A few years later, Geofferey Masters (1982) added the Partial Credit model to the family. All these Rasch models are based on the idea that useful measurement involves the examination of only one human attribute at a time on some hierarchy of “less than / more than” on a single continuum of interest (e.g., attitude/opinion) (Bond & Fox, 2001). Even if a variable has many characteristics, only one of them can be meaningfully rated at a time. A variable is thus conceptualized as a continuum of “less than / more than” of each of these characteristics. In other words, the model postulates that the underlying trait being measured (unique dimension) can entirely accounts for the responses gathered, and each item is considered as an indirect measure of this trait (Martin, Campanelli, & Fay, 1991).

These Rasch models also assume that the respondents’ answers to the items are statistically independent, i.e. that each answer is only determined by the joint effect of the respondent’s parameter and the item’s parameter. A person’s parameter is thus assumed to reflect each respondent’s value along the continuum of the variable being measured. Likewise, an item’s parameter is assumed to reflect the position of the characteristic of the variable along that same continuum. And the odds of a person agreeing with each item, is the product of an item parameter and a person parameter. This is what is referred to as the separability of item and person parameters. Finally, items’ parameters are

assumed not to vary over respondents and persons' parameters are assumed not to depend on which question is being asked (Martin et al., 1991). This is what is referred to as the property of invariance.

To estimate the person and item parameters, these models use a probabilistic form of the Guttman scale (Keenan, Redmond, Horton, Conaghan, & Tennant, 2007) that shows what should be expected in the response patterns of the items and against which they are tested. These models consider that all persons are more likely to endorse items that are easy to agree with than to endorse items that are difficult to endorse. Likewise, all items are more likely to be endorsed by persons of high "agreeability" than by persons of low "agreeability". As a result, if a person has agreed to an item of an average level of endorsement toward something, then all items below that level of endorsement should also be endorsed by that person. On the other hand, any item over that level of endorsement should be harder to endorse by that person. Similarly, if an item has been endorsed by a person of an average level of "agreeability", than it should also be endorsed by persons of higher level of agreeability. However, it should not be endorsed by persons of lower level of agreeability.

Mathematically, these three Rasch models assume that the probability P_{ni} of a person n, endorsing (or agreeing) with an item i, can be formulated as a logistic function of the relative distance between the item location D_i (the position of the characteristic of the attitude/opinion being measured as expressed by this item) and the person location B_n (the level of agreeability of this person toward the attitude/opinion being measured) on a linear scale (the continuum of "less than / more than" of the attitude/opinion being measured). As a result, both the item and the person parameters are presented on the same log-odds units (logit) scale.

The mathematical expression of the dichotomous Rasch model (Rasch, 1960/1980) is thus :

$$P_{ni} = \frac{e^{(Bn - Di)}}{1 + e^{(Bn - Di)}}$$

Taking the natural logarithm of the odds ratio, the expression becomes a logit model:

$$Ln\ [\ P_{ni} / (1 - P_{ni})\ ] = B_n - D_i$$

To extend this model to the polytomous case, namely the Rating Scale model, another parameter must be introduced. Data obtained from Likert-type rating scales are usually analyzed as if the response options were equal-interval (J. Dawes, 2007). The Rasch rating scale model does not presume that the size of the step between each category is equal. Instead, the model analyses the data and establishes the pattern in the use of the scale categories. It can then produce a rating scale structure shared by all items (Bond & Fox, 2001). Rasch modelling transforms the counts of endorsement in each response category of the rating scale into an interval scale based on the actual data. As a result, in addition to the person and item parameters, the model also estimates a series of thresholds for the scale used. These thresholds are the level at which the likelihood of non endorsement at a given response category (below the threshold) turns to the likelihood of endorsement at that category (above the threshold). As an example, one of the threshold of a rating scale using four response options with the semantic labelling “disagree totally”, “disagree”, “agree”, “agree totally”, would be located between the options “disagree totally” and “disagree”, at the position where a respondent would fail to endorse the “disagree totally” but endorse the “disagree” option.

As opposed to the Rating Scale model which considers that the threshold structure is the same across all items, the Partial Credit model allows the threshold structure to vary across items. Consequently, a simple formulation of the Partial Credit model (Masters, 1982) is :

$$\frac{e^{(Bn - Di - Fix)}}{1 + e^{(Bn - Di - Fix)}}$$

Taking the natural logarithm of the odds ratio, the expression becomes :

$$Ln\ [\ P_{nix} / P_{ni(x-1)}\ ] = B_n - D_i - F_{ix}$$

Where P_{nix} is the probability that person n with attitude/opinion B_n endorses category x (where x = 0 to m-1 for the m response categories of the rating scale) of item i located at position D_i on the variable continuum. Parameter F_{ix} corresponds to the threshold between categories x-1 and x on item i; or, more precisely, to the point at which the probability of opting for one or the other category on item i is equal. F_{ix} can also be interpreted as the distance between category x-1 and category x on item i. It is through this parameter that the model can accommodate for a different number of categories x for

each item *i*. (In the Rating Scale model, this parameter would simply be F_x because the number of categories *x* and threshold structure is the same for all items.) Finally, $P_{ni(x-1)}$ represents the probability that person n with opinion B_n endorses category *x*-1 on item *i*.

3.2 Quality of fit between the data and the model

As mentioned in the previous section, the Partial Credit model uses a probabilistic form of the Guttman scale to show what should be expected in the response patterns of the items and against which the data are tested. This means that before using the modeled data, there is a calibration phase that is required, in which the observed data are tested against the model to evaluate the goodness-of-fit. Because the model defines what are appropriate measurement values, data must first meet the model's expectation and not the contrary (Pallant & Tennant, 2007). This is in opposition to the usual "statistical" view where models are developed to best represent the data. From a statistical point of view, the model can be considered as a null hypothesis: if the goodness-of-fit tests yield to significant results, the null hypothesis has to be rejected and the model is not a valid model for the data at hand (Verhelst & Glas, 1995).

The software that was used in our analysis, RUMM2020, uses three overall fit statistics to determine if the data fit the model or not. Two of them are item-person interaction statistics. Their value is the standardized sum of all differences between observed and expected values summed over all persons (persons fit residual) and over all items (items fit residual). Because they are standardized, a perfect fit to the model for the persons or the items would give a mean of zero and a standard deviation of 1. The third fit statistic is an item-trait interaction, reported as a chi-square, and reflecting the property of invariance across the trait. For each item, RUMM2020 calculates a chi-square statistic that compares the difference between observed values and expected values across groups representing different levels of ability (called class intervals) along the continuum of the trait being measured. Therefore, for a given item, several chi-square values are summed to give the overall chi-square for the item, with degrees of freedom being the number of groups minus 1 (Tennant & Conaghan, 2007). The item-trait interaction chi-square statistic is thus the sum of the chi-squares for individual items across all items. Bonferroni corrections are applied to adjust the *p* value of these chi-square statistics to take into

account the multiple values computed. A significant chi-square indicates that the hierarchical ordering of the items varies across the trait, thus compromising the required property of invariance.

In addition to these statistics, RUMM2020 reports individual person and item fit statistics. The value of these indices are the standardized sum of all differences between observed and expected values summed over a person (individual person fit residual) or an item (individual item fit residual). The chi-square statistic of each item is also reported. Therefore, when a person or a group of persons do not fit the model, it is possible to remove them from the sample. The same would apply to misfit items. The difference between these two actions is mainly a question of quantity since in studies using a questionnaire there are generally more persons than items included. Moreover, inclusion of an item in a questionnaire is generally done for reasons related to validity and eliminating an item on pure statistical grounds may afflict the validity of the data gathered for the measurement of a given construct (Verhelst & Glas, 1995). Also, development of items in a professional setting may be quite expensive. As a result, it is often easier to exclude persons from the analysis than to exclude items. However, eliminating persons is not without consequences with regard to the generalizability of the results.

Other tools are also available in RUMM2020 to help one investigate the goodness-of-fit between the data and the model. First, there is the person separation index. This statistic, like the traditional reliability, depends in part on the actual variance of the persons[1]. Very similar to Cronbach's alpha, it is estimated as the ratio of the true to the observed variance. It's interpretation is also done in a similar manner: a minimum value of 0.7 is recommended for group use and 0.85 for individual use (Tennant & Conaghan, 2007). The person separation index is an indicator of the degree to which the relative variation amongst the persons is not random variation[2]. The category probability curves, as well as the threshold probability curves, of each item can also be inspected. The model considers that all persons are more likely to endorse items that are easy to agree with than to endorse items that are difficult to agree with. Likewise, it considers that all items are more likely to be endorsed by persons of high "agreeability" than by persons of low "agreeability". Therefore, one would expect that, if the data fits the model, each response option would systematically take turn in showing the highest probability of

[1] RUMM Laboratory (2004). Interpreting RUMM2020 Part I: Dichotomous Data. p. 9.
[2] RUMM Laboratory (2005). Interpreting RUMM2020 Part II: Polytomous Data. p. 35.

endorsement along of the continuum of the trait being measured (Tennant & Conaghan, 2007). When respondents fail to use the response categories in a manner consistent with what is expected by the model, i.e. when respondents have difficulty discriminating between the response categories or when the labelling of the options is potentially too confusing, occurs what is referred to as “disordered thresholds”. This is one of the most common sources of item misfit. In such situations, it is possible to collapse the categories where disordered thresholds occur. Often, it will improve the overall fit to the model. But which categories should be collapsed?

3.3 Disordered thresholds and collapsing categories

Disordered thresholds indicate a failure to construct a measure from the response scale provided, i.e. from the ordered categories offered, represented by successive scores, and supposed to reflect an increasing level of the latent trait (attitude or opinion) being measured. As an example, consider a response scale consisting of three ordered categories : « disagree », « neutral » and « agree ». Consider also person A, whose level of « agreeability » for an attitude or an opinion is at the threshold between the « disagree » and the « neutral » categories. Consider that another person B, would have a level of « agreeability » located at the threshold between the « neutral » and the « agree » categories. Clearly, person B has a higher level of « agreeability » than person A. But, disordered threshold implies that the level of « agreeability » estimated for person A is higher than the one estimated for person B. In other words, the estimates provided by the model indicate that the manner in which the opinion/attitude is being measured is in opposition to the manner in which it was intended. As a result, the estimates provided cannot be taken as they are.

When items show disordered thresholds, it is possible to collapse categories until the thresholds are properly ordered or until items show adequate fit to the model (Tennant, 2004). There is no unique way of collapsing categories. Bond and Fox (2001, p. 167) propose some guidelines for collapsing categories. One of their guideline is that collapsing two categories must make sense. Therefore, before collapsing categories that show disordered threshold, one should wonder if the combination of, let say, the « disagree » and the « neutral » categories makes sense for the attitude/opinion being measured. Linacre (2004) also suggested a few guiding tips to help optimize a rating scale’s effectiveness. First, Linacre mentions that there should be about 10 observations or more in each

category of the scale. Also, he indicates that the observations should be uniformly distributed in each category to obtain an optimal calibration of the scale. As a last example, Tennant (2005) suggests to look at the person separation index as a guide, as well as the fit statistics (e.g. the chi-square interaction). The solution that gives the highest personal separation index, all things being equal and given fit to the model, is the solution providing the greatest precision. With so many different "guidelines", we decided to explore these and many others to find which one would bring our data to fit the model optimally.

4. Method

4.1 The instrument and the respondents

The instrument is a self-administered survey questionnaire that was developed by the Centre de Formation Initiale des Maîtres (CFIM) of the Université de Montréal in 1999 to gather data for the assessment of its undergraduate teacher-training program. The questionnaire is written in French and was first distributed during the spring of 2000. Data was collected every spring since that time and until 2007, but using two versions of the same questionnaire.

The original version of the questionnaire (used in 2000) was made up of eight sections. Throughout the years, the questionnaire was modified so that in the 2007 version, only four sections remained: overview of the training, teacher training, internships, various information of a demographic nature. Only the teacher training section was retained for the research carried out. In this section, students must respond to 24 items introduced by the prompt line "I consider that my program has enabled me to develop competencies for…".

Participants were offered a bipolar rating scale to record their answers. In version A of the questionnaire, the scale was made of five response categories with semantic: "1 = Disagree totally", "2 = Rather disagree", "3 = Neutral", "4 = Rather agree" and "5 = Agree totally". In version B of the questionnaire, the scale was made of six response categories with semantic: "1 = Disagree totally", "2 = Mainly disagree", "3 = Somewhat disagree", "4 = Somewhat agree", "5 = Mainly agree" and "6 = Agree totally". Since the optimal number of response options on a scale tends to be close to seven

response categories (according to what was presented in previous sections), version B, with six response categories, was the one retained for our analysis.

The two versions of the questionnaire were distributed during the spring of 2007. Since the questionnaire is used to evaluate undergraduate teacher training programs, there is no sampling of individuals. The intended respondents are all fourth-year students from the teacher training program for preschool and elementary school at the Université de Montréal. Each pile of questionnaires that was distributed to the students was an alternation of a version B following a version A. The questionnaires were then distributed randomly to the students at the end of their last semester through regular courses.

4.2 Data processing software

Modelling polytomous scales like the ones used in this paper requires quite complex processes of calculation. Drawing the characteristics curves of items, estimating parameters or verifying the basic hypothesis associated with the models necessitate the use of specialized software. Many softwares enabling one to apply Rasch models are available on the market. Among them, let us cite Bilog, Conquest, Winsteps or Rumm2020. Rumm2020 has useful features for the kind of study proposed in this paper. It is therefore the software that was retained to analyse our data.

5. Data analysis

CFIM's undergraduate teacher-training program assessment for the year 2007, yield to the gathering of responses from 117 students from the preschool and elementary school program. Sixty of these students completed version A of the questionnaire and 57 completed version B. Since only version B of the questionnaire was retained for our analysis, our sample was composed of 57 students.

Many strategies to collapse categories were explored in this study. Table 1 summarizes the results obtained. (Tables and figures are presented in annex 1 at the end of the paper.) To interpret the results, it should be noted that, as a general rule, if an estimate converge quickly, it is a good sign that the data are in accord with the model. However, a large number of categories inevitably require a lot more iterations to converge. Therefore, to determine the best strategy within the ones we tested, we

will look for the largest number of parameters who converged after 100 iterations. Also, we mentioned in section 3.3 that disordered thresholds are an indication of a failure to construct a measure from the response scale provided on the questionnaire or, in other words, that the data collected is not congruent with the expected model. Consequently, the best strategy will also be the one who presents the fewest number of items with disordered thresholds. Moreover, Rumm2020 provides an item-trait interaction chi-square statistic. A significant chi-square indicates that the hierarchical ordering of the items varies across the trait, thus compromising the required property of invariance (section 3.3). This is another aspect we will take into account to determine the best strategy. The alpha value used to determine if a value is statistically significant is fixed at 0.05, since the purpose of our analysis does not require this value to be smaller. Finally, acording to Tennant (2005), the solution that gives the highest personal separation index, all things being equal, and given fit to the model, is the solution providing the greatest precision (section 3.3). As a result, the best strategy will be the one with the fewest number of misfit items and misfit persons, but with the highest person separation index. According to Lawton, Bhakta, Chamberlain and Tennant (2004), fit residual statistic should be included in the interval [-2.5, 2.5]. In our analysis, since our sample is quite small, we decided to extend this interval to [-3, 3] in order to keep as much items and persons as possible.

A first analysis was done on all subjects and all items. The scoring structure used for this analysis corresponds to the response scale presented on version B of the questionnaire. Rumm2020 indicates that 69 parameters out of 96 converged after 100 iterations. In this analysis, 13 items show disordered thresholds. The items fit residual is 0.4 (S.D. = 1.1) and the persons fit residual is -0.5 (S.D. = 2.3). According to these statistics, items show a better fit to the model than persons as their fit residual is closer to zero and standard deviation is closer to 1. The item-trait interaction statistic has a significant chi-square *p* value of 0.001 indicating that the hierarchical ordering of the items varies across the trait. The person separation index is 0.94. Analysis of individual item fit residuals shows that none of the items is outside the fit values. One item (item 23) has a significant chi-square *p* value of 0.0002, which is below the Bonferroni adjustment. Analysis of individual person fit residuals indicates that there are 11 misfit persons. In light of these results, we find that the Rating Scale model is not a good model for the data at hand.

In a second analysis, misfit persons were withdrawn from the sample to see if a better fit to the model would be obtained. After an iterative process, a total of 13 persons were removed from our sample. Results show that, again, only 69 out of 96 parameters converged after 100 iterations. Twelve items show disordered thresholds. The items fit residual is 0.4 (S.D. = 0.9) and the persons fit residual is 0.001 (S.D. = 1.5). The item-trait interaction statistic is not significant meaning that the hierarchical ordering of the items does not vary across the trait. The person separation index is 0.94. None of the items is outside the chosen fit interval, but item 23 still has a significant chi-square *p* value with the Bonferroni adjustments (*p* =0.0005). Therefore, it seems that the withdrawal of misfit persons, without collapsing any categories, doesn't improve much the results. Our next analysis will thus focus on collapsing categories to see if it allows us to obtain better results.

The study of the category probability curves of these two analysis revealed that for most of the items, categories 2 and 3[3] never had more chances than the other categories to be chosen (see figure 1 as an example). This means that these categories were probably indistinguishable for people and that, even if they were offered on our questionnaire, there is no actual threshold or boundary between these two categories (Andrich, 1996). Therefore, as a third analysis, we decided to try and collapse these two categories, using all items and all persons, to create some sort of midpoint or ambivalent category. Results reveal that only 53 parameters converged after 100 iterations. Twelve items still show disordered thresholds and not necessarily the same twelve items as in previous analysis. The item-trait interaction is significant (chi-square *p* value of 0.002). The person separation index is 0.94. None of the items is outside the chosen interval but item 5 has a significant chi-square *p* value with the Bonferroni adjustments (p = 0.0008). Thirteen persons have their individual person fit residual outside of the chosen interval.

Since the results obtained in this third analysis were quite similar to the first ones, we decided to apply these collapsed categories on the sample where the misfit persons were removed (fourth analysis). Again, it does not improve much the results. (See table 1 for more details.) Data still do not fit well with the model.

[3] It should be noted that in Rumm2020, responses categories must be rescored from 0 to 5. As a result, categories 2 and 3 on figure 1 correspond to categories 3 and 4 on our questionnaire.

We also tried to collapse category 2 with the collapsed categories 3 and 4 on this reduced sample, since the threshold between category 2 and category 3 (threshold 2) appeared problematic (see figure 2 as an example). This is analysis n. 5. Only 36 parameters converged after 100 iterations. On the other hand, only 3 items show disordered thresholds. This is a clear improvement. The item-trait interaction is not significant. The person separation index is 0.92. One item (item 23) is outside the chosen fit interval. None of the items has a significant chi-square *p* value. Only 2 persons have their individual person fit residual outside of the chosen interval. This solution does seem to improve how data fit to the model. However, we think that collapsing the “Mainly disagree”, the “Somewhat disagree” and the “Somewhat agree” categories to create some sort of large “neutral” category may cause conceptual problems. As Bond and Fox mentioned (2001), collapsing categories must make sense.

Our next attempt was thus to collapse the intermediate categories. Indeed, collapsing the “Mainly disagree” and the “Somewhat disagree” categories does make more sense. As a first step, we started by collapsing categories 2 and 3 (analysis n. 6). Only 60 parameters converged after 100 iterations. One item show disordered thresholds. The item-trait interaction is significant (chi-square *p* value of 0.02). The person separation index is 0.94. None of the items is outside the chosen interval but item 23 has a significant chi-square *p* value with the Bonferroni adjustments (p = 0.0002). Thirteen persons have their individual person fit residual outside of the chosen interval. Again, we see an improvement as to the number of items showing disordered thresholds, but the number of misfit persons is still high.

The next step was to collapse categories 2 and 3, as well as categories 4 and 5 (analysis n. 7). Only 36 parameters converged after 100 iterations. All items show ordered thresholds. The item-trait interaction is not significant. The person separation index is 0.92. None of the items is outside the chosen fit interval but item 2 has a significant chi-square *p* value with the Bonferroni adjustments (*p* = 0.0002). Nine persons have their individual person fit residual outside of the chosen interval. Again, we find a little improvement in the results: now all items show ordered thresholds and the number of misfit persons is a little less.

Analysis of the categories frequencies revealed that 6 items had null frequencies in their first category (items 1, 7, 8, 22, 23 and 24). As a result, we tried to rescore these 6 items by combining category 1 to the collapsed categories 2 and 3, keeping categories 4 and 5 collapsed too (analysis n. 8). All parameters converged after 100 iterations. All items show ordered thresholds. The item-trait interaction is significant (chi-square *p* value of 0.03). The person separation index is 0.92. None of the items is outside the chosen fit interval but item 2 has a significant chi-square *p* value with the Bonferroni adjustments ($p = 0.0002$). Nine persons have their individual person fit residual outside of the chosen interval. Once more, we find improvements in our results since all parameters converged. So far, collapsing intermediate categories, which makes sense, provides good results. Moreover, it seems that a solution that applies to each item separately may work better than a solution that applies equally to all items. In other words, the Partial Credit model appears to be a better model for the data at hand.

In order to investigate other collapsing strategies we then referred to Cools *et al.* (2006) who found that supplementary extreme categories such as “fully agree” or “fully disagree” did not improve the metric properties of their scale. Our next attempt thus consisted in collapsing categories 1 and 2 (analysis n. 9). Results indicate that 93 parameters converged after 100 iterations. Seventeen items show disordered thresholds. The item-trait interaction is significant (chi-square *p* value of 0.002). The person separation index is 0.94. None of the items is outside the chosen interval but item 23 has a significant chi-square *p* value with the Bonferroni adjustments ($p = 0.0003$). Thirteen persons have their individual person fit residual outside of the chosen interval.

We then tried to collapse categories 5 and 6 (analysis n. 10). Results show that only 53 parameters converged after 100 iterations. Twenty-one items show disordered thresholds. The item-trait interaction is not significant. The person separation index is 0.94. None of the items is outside the chosen interval and none has a significant chi-square *p* value. Six persons have their individual person fit residual outside of the chosen interval, and two have extreme fit residual values.

Collapsing categories 1 and 2, as well as categories 5 and 6, doesn’t give much better results (analysis n. 11). Only 46 parameters converged after 100 iterations. Sixteen items show disordered

thresholds. The item-trait interaction is not significant. The person separation index is 0.94. None of the items is outside the chosen interval and none has a significant chi-square *p* value. Again, six persons have their individual person fit residual outside of the chosen interval, and two have extreme fit residual values. Consequently, Cools *et al.*' suggestion does not help us since the data do not fit well with the model.

Our next attempts intended to verify Linacre's suggestions. Therefore, we first tried to combine categories to reach, as much as possible, a minimum of 10 responses in each category (analysis n. 12). All parameters converged after 100 iterations. Eleven items show disordered thresholds. The item-trait interaction is significant (chi-square *p* value of 0.002). The person separation index is 0.94. None of the items is outside the chosen fit interval but item 23 has a significant chi-square *p* value with the Bonferroni adjustments ($p = 0.0004$). Ten persons have their individual person fit residual outside of the chosen interval. In light of these results, it seems that this suggestion may apply more to analysis done with Winsteps than with Rumm2020. Indeed, the conditional pairwise estimation procedure used by Rumm2020 estimates threshold parameters from all data, and not just from adjacent categories like in Winsteps, enhancing the stability of estimates.

Our second analysis consisted in combining categories to obtain, as much as possible, a uniform, and unimodal, distribution of frequencies across the different categories (analysis n. 13). Results show that all parameters converged after 100 iterations. Only one item show disordered thresholds. The item-trait interaction is significant (chi-square *p* value of 0.007). The person separation index is 0.94. None of the items is outside the chosen fit interval but item 23 has a significant chi-square *p* value with the Bonferroni adjustments ($p = 0.0005$). Nine persons have their individual person fit residual outside of the chosen interval. As a result, it seems that having a uniform distribution does help improve the results, but they do not provide the best results of our analysis. However, once more, it shows that solutions that applies specifically to each item instead of a general solution applied to all items seems preferable and provides better fit to the model.

In sum, a total of 13 strategies were tested. Among them, only 3 allowed all parameters to converge (8, 12 and 13). Three strategies minimized the number of items with disordered thresholds (5, 6 and

13) and two corrected the problem for all items (7 and 8). The item-trait interaction was not significant for 6 strategies (2, 4, 5, 7, 10 and 11). The person separation index was the highest for analysis number 1 and 9, although it did not vary much through the different strategies tested. Overall, items showed good fit to the model, except for strategy number 5 where one item was identified as misfit. For most of the analysis, one item had a significant chi-square, except for strategies number 4, 5, 10 and 11 where none of the chi-square was significant. Finally, unless misfit persons were removed from the sample, all strategies identified misfit persons. On the complete sample, three analysis minimised the number of misfit persons (7, 8 and 13), although this number did not vary much through the different strategies tested.

6. Discussion and conclusion

Analysis done in the previous section illustrated how different methods used to collapse categories can provide quite different results. First, collapsing the mid-scale categories (*i.e.* categories 2, 3 and 4) provided interesting results. However, collapsing “Mainly disagree” with “Somewhat disagree” and “Somewhat agree” may cause conceptual problems.

In a similar way, Linacre’s suggestion to collapse categories in order to obtain a uniform distribution, did seem to help improve the quality of fit between the data and the model, but did not provide the best results. To reach a uniform distribution, we had to collapse categories 1, 2 and 3 for most of the items. This means that, for these items, all the disagree categories were combined while the agree options remained. This causes an imbalance in a scale that was intended to be bipolar. Moreover, it causes an important lost of information as to the real level of disagreement the respondents have in regard to the different items of the questionnaire. In some contexts, such a lost might have an impact on the conclusions drawn from the research. Therefore, although this solution provides interesting results, it should be applied cautiously.

Collapsing the intermediate categories (somewhat and mainly) was a strategy that provided among the best results. These results were even better when, in addition to combining the intermediate categories, we also tried to avoid null frequencies. As a result, we think that collapsing categories must, first and foremost, make sense. Then, other strategies, like Linacre’ suggestions may help

improve the results obtained. Also, we found that, most of the times, general solutions applied equally to all items provided poorer results than solutions applied specifically for each item. This tells us that when collapsing categories, maybe it's not one size fits all.

Finally, looking at the person separation index does not seem to help very much. Indeed, in all our analysis, the person separation index value remained almost constant. Moreover, it's value was even lower when data better fitted the model, then when the strategy used to collapse categories provided poor quality of fit. In a similar way, we found that the item-trait interaction was not as helpful as we would have thought. This statistic was generally significant when a strategy provided good quality of fit between the data and the model, and not significant when the quality of fit was poorer. On the other hand, we found that the number of parameters who converged, the number of items with disordered thresholds, the number of misfit items and misfit persons were helpful tools.

It should be noted that this research is an exploratory study and that the size of the sample is limited. Consequently, the results obtained could be unstable and lacking in precision. Therefore, other researches would be necessary to explore these strategies in other contexts and to confirm the results obtained here.

As a general conclusion, we found that collapsing categories is not necessarily intuitive. Although many guidelines exist to help one make a decision on how categories should be collapsed, they remain guidelines that should not be applied blindly, but that should be adapted to each context.

References

Alwin, D. F., & Krosnick, J. A. (1991). The reliability of survey attitude measurement: The influence of question and respondent attributes. *Sociological Methods & Research, 20*(1), 139-181.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*(4), 561-573.

Andrich, D. (1996). Category ordering and their utility. *Rasch Measurement Transactions, 9*(4), 464-465.

Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model : Fundamental measurement in the human sciences*. Mahwah, N.J.: Lawrence Erlbaum Associates.

Borg, G. (2001). Are we subjected to a 'long-standing measurement oversight'? *Proceedings of Fechner Day 2001*, The International Society of Pshychophysics. Retrieved from www.ispsychophysics.org/component/option,com_docman/task,cat_view/gid,4/Itemid,38/.

Chang, L. (1994). A psychometric evaluation of 4-point and 6-point Likert-type scales in relation to reliability and validity. *Applied Psychological Measurement, 18*(18), 3.

Cook, K. F., Amtmann, D., & Cella, D. (2006, 6-11 avril). *Is more less? Impact of number of response categories in self-reported pain.* Paper presented at the annual meeting of the American Educational Research Association (AERA), San Francisco, CA.

Cools, W., Hofmans, J., & Theuns, P. (2006). Context in category scales: Is "fully agree" equal to twice agree? *Revue Européenne de Psychologie Appliquée, 56*, 223-229.

Cox, E. P. (1980). The optimal number of response alternatives for a scale: A review. *Journal of Marketing Research, 17*(4), 407-422.

Davies, R. S. (2008). Designing a response scale to improve average group response reliability. *Evaluation and Research in Education, 21*(2), 134-146.

Dawes, J. (2007). Do data characteristics change according to the number of scale points used? An experiment using 5-point, 7-point and 10-point scales. *International Journal of Market Research, 50*(1), 61-77.

Dawes, R. M., & Smith, T. L. (1985). Attitude and opinion measurement. In G. Lindzey & E. Aronson (Eds.), *Handbook of social psychology* (Third ed., Vol. 1: Theory and method, pp. 509-566). New York: Random House.

Hawthorne, G., Mouthaan, J., Forbes, D., & Novaco, R. W. (2006). Response categories and anger measurement: Do fewer categories result in poorer measurement? *Social Psychiatry & Psychiatric Epidemiology, 41*, 164-172.

Keenan, A.-M., Redmond, A. C., Horton, M., Conaghan, P. G., & Tennant, A. (2007). The foot posture index: Rasch analysis of a novel, foot-specific outcome measure. *Archives of Physical Medecine and Rehabilitation, 88*, 88-93.

Kirnan, J. P., Edler, E., & Carpenter, A. (2007). Effect of the range of response options on answer to biographical inventory items. *International Journal of Testing, 7*(1), 27-38.

Klopfer, F. J., & Madden, T. M. (1980). The middlemost choice on attitude items: ambivalence, neutrality or uncertainty? *Personality and Social Psychology Bulletin, 6*(1), 97-101.

Krosnick, J. A., & Fabrigar, L. R. (1997). Designing rating scales for effective measurement in surveys. In L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz & D. Trewin (Eds.), *Survey measurement and process quality* (pp. 141-164). New York: John Wiley & Sons, Inc.

Lawton, G., Bhakta, B. B., Chamberlain, M. A., & Tennant, A. (2004). The Behçet's disease activity index. *Rheumatology, 43*(1), 73-78.

Likert, R. (1932). *A technique for the measurement of attitudes.* Archives of psychology, No 140. New York: R. S. Woodworth.

Linacre, J. M. (2004). Optimizing rating scale category effectiveness. In E. V. Smith Jr. & R. M. Smith (Eds.), *Introduction to Rasch measurement: Theory, models and applications* (pp. 258-278). Maple Grove, MN: JAM Press.

Martin, E. A., Campanelli, P. C., & Fay, R. E. (1991). An application of Rasch analysis to questionnaire design: Using vignette to study the meaning of 'Work' in the current population survey. *The Statistician (Special issue: Survey design, methodology and analysis (2)), 40*(3), 265-276.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *The Psychological Review, 63*(2), 81-97.

Pallant, J. F., & Tennant, A. (2007). An introduction to the Rasch measurement model: An example using the Hospital Anxiety and Depression Scale (HADS). *British Journal of Clinical Psychology, 46*, 1-18.

Poulton, E. C. (1989). *Bias in quantifying judgments*. Hove, UK: Lawrence Erlbaum Associates.

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests*. Chicago: The University of Chicago Press.

Schaeffer, N. C., & Presser, S. (2003). The science of asking questions. *Annual review of sociology, 29*, 65-88.

Tennant, A. (2004). Disordered Thresholds: An example from the Functional Independence Measure. *Rasch Measurement Transactions, 17*(4), 945-948.

Tennant, A. (2005). [MBC-Rasch] what to do? (Online publication. Retrieved July 2009, from Rasch mailing list: https://lists.wu-wien.ac.at/pipermail/rasch/2005q1/000352.html

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 & Rheumatism, 57*(8), 1358-1362.

Verhelst, N. D., & Glas, C. A. W. (1995). The one parameter logistic model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 215-237). New York: Springer-Verlag.

Weathers, D., Sharma, S., & Niedrich, R. W. (2005). The impact of number of scale points, dispositional factors, and the status quo decision heuristic on scale reliability and response accuracy. *Journal of Business Research, 58*, 1516-1524.

Weng, L.-J. (2004). Impact of the number of response categories and anchor labels on coefficient alpha and test-retest. *Educational and Psychological Measurement, 64*(6), 956-972.

Wright, B., & Linacre, J. (1992). Combining and splitting categories. *Rasch Measurement Transactions, 6*(3), 233-235.

Annex 1

Table 1 : Summary of the different strategies explored to collapse categories

Description of the analysis	Number of parameters who converged after 100 iterations	Number of items with disordered thresholds	Item-trait intereraction (chi-square *p* value)[4]	Person separation index	Number of misfit items[5]	Individual item fit (chi-square *p* value)	Number of misfit persons
1. Initial analysis: all items, all persons, scale structure as presented on the questionnaire.	69	13	0.001119*	0.94134	None	Item 23 (*p*= 0.000179)	11
2. All items, scale structure as on the questionnaire, misfit persons removed from the sample .	69	12	0.093235	0.93582	None	Item 23 (*p*= 0.000509)	None
3. All items, all persons, mid-scale categories collapsed (*i.e.* categories 3 and 4).	53	12	0.002490*	0.93529	None	Item 5(*p*= 0.000750)	13
4. All items, misfit persons removed, mid-scale categories collapsed (*i.e.* 3 and 4).	53	12	0.077671	0.92994	None	None	1
5. All items, misfit persons removed, mid-scale categories collapsed (*i.e.* 2, 3 and 4).	36	3	0.276752	0.91678	Item 23	None	2
6. All items, all persons, intermediate categories collapsed (*i.e.* 2 and 3).	60	1	0.019861*	0.93897	None	Item 23 (*p*= 0.000165)	13
7. All items, all persons, intermediate categories collapsed (*i.e.* 2-3 and 4-5).	36	None	0.061847	0.92213	None	Item 2 (*p*= 0.000224)	9
8. Same as analysis n. 7, but for 6 items, the collapsed categories are 1-2-3 and 4-5 to avoid null frequencies.	All	None	0.002722*	0.92219	None	Item 2 (*p*= 0.000223)	9
9. All items, all persons, extreme categories collapsed (*i.e.* 1 and 2).	93	17	0.001920*	0.94002	None	Item 23 (*p*= 0.000339)	13
10. All items, all persons, extreme categories collapsed (*i.e.* 5 and 6).	53	21	0.450169	0.93842	None	None	6 + 2 extremes
11. All items, all persons, extreme categories collapsed (*i.e.* 1-2 and 5-6).	46	16	0.140354	0.93738	None	None	6 + 2 extremes
12. For each item separately, categories were collapsed to reach, as much as possible, a minimum of 10 observations.	All	11	0.001581*	0.93619	None	Item 23 (*p*= 0.000362)	10
13. For each item separately, categories were collapsed to obtain a uniform distribution.	All	1	0.007466*	0.93623	None	Item 23 (*p*= 0.000504)	9

[4] A chi-square *p* value is considered significant (*) when it's value is below the alpha value of 0.05.
[5] Misfit items are items for which the individual fit statistic is outside the interval [-3, 3].

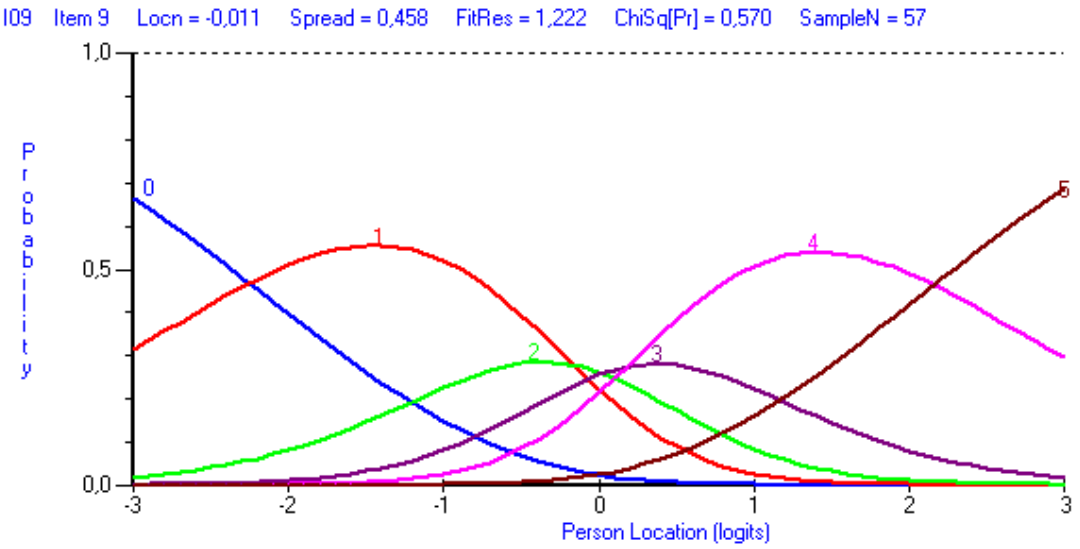

Figure 1 : Category probability curve for item 9, when all subjects and all items are included, and the scoring structure is as shown on version B of the questionnaire.

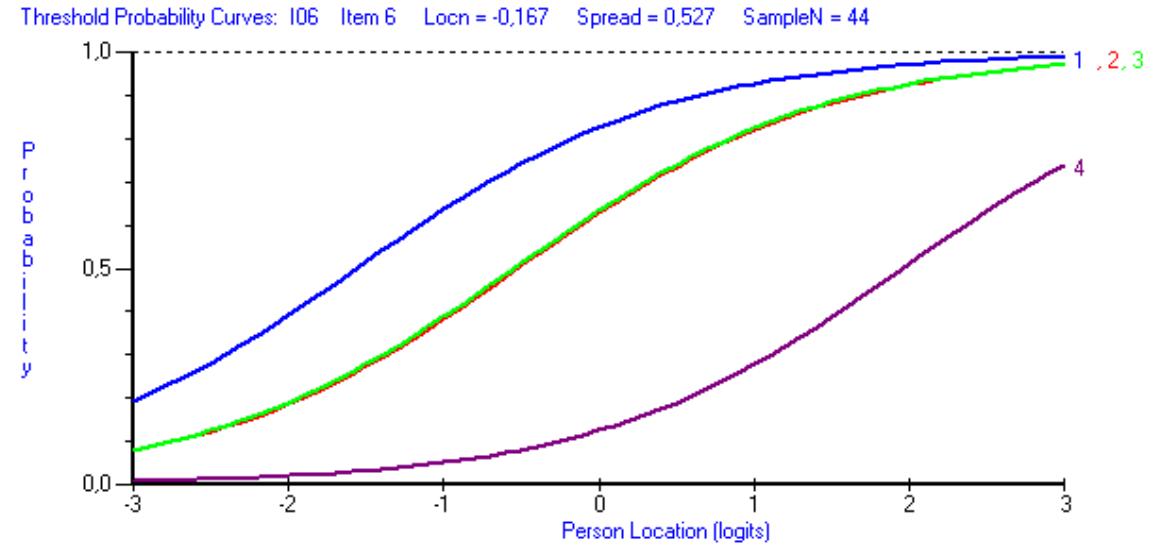

Figure 2 : Threshold probability curves for item 6, when misfit persons are removed from the sample, and categories 3 and 4 are collapsed.