Received July 2005
Revised November 2005
Accepted December 2005

Reconsidering the problem of data equivalence in international marketing research

Contrasting approaches based on CFA and the Rasch model for measurement

Thomas Salzberger
Vienna University of Economics and Business Administration, Vienna, Austria, and
Rudolf R. Sinkovics
Manchester Business School, UK

Abstract

Purpose – The paper investigates the suitability of the Rasch model for establishing data equivalence. The results based on a real data set are contrasted with findings from standard procedures based on CFA methods.

Design/methodology/approach – Sinkovics *et al.*'s data on technophobia was used and re-evaluated using both classical test theory (CTT) (multiple-group structural equations modelling) and Rasch measurement theory.

Findings – Data equivalence in particular and measurement in general cannot be addressed without reference to theory. While both procedures can be considered best practice approaches within their respective theoretical foundation of measurement, the Rasch model provides some theoretical virtues. Measurement derived from data that fit the Rasch model seems to be approximated by classical procedures reasonably well. However, the reverse is not necessarily true.

Practical implications – The more widespread application of Rasch models would lead to a stronger justification of measurement, in particular, in cross-cultural studies but also whenever measures of individual respondents are of interest.

Originality/value – Measurement models outside the framework of CTT are still scarce exceptions in marketing research.

Keywords Data analysis, Set theory, Measurement, Cross-cultural studies

Paper type Research paper

1. Introduction and purpose

Advances in International Marketing research rely on cross-country comparisons and require sound conceptualisations and empirical generalisations. Meaningful comparisons are possible only if data are derived from equivalent constructs and associated measures. Issues of reliability, validity as well as dimensional consistency need to be addressed (Davis *et al.*, 1981; van de Vijer and Leung, 1997) in order to avoid misinterpretation of results and fallacious conclusions for managerial decision making.

On the one hand, advancements in statistical and computational methodologies and procedures (Tabachnick and Fidell, 2001) have helped solve methodological problems. On the other hand, applied researchers and practitioners are faced with an accumulation

International Marketing Review
Vol. 23 No. 4, 2006
pp. 390-417

0265-1335
DOI 10.1108/02651330610678976

of competing frameworks and methods for analysing cross-national datasets (van de Vijer and Leung, 1997). Scholars have looked at measurement equivalence issues from both a qualitative and quantitative perspective (Mullen, 1995; Salzberger *et al.*, 1999; Singh, 1995). There is also a heated discussion around scale development frameworks and the nature of operationalisations, incited by Rossiter, 2002 C-OAR-SE paradigm (Diamantopoulos, 2005; Rossiter, 2005). Empirically, multi-group structural equations modelling (Steenkamp and Baumgartner, 1998), which is grounded in classical test theory (CTT), and to a lesser degree generalizability theory as an extension of CTT (Finn and Ujwal, 2005; Rentz, 1987) have gained popularity.

The main purpose of this paper is to examine the use and applicability of an alternative measurement approach, Rasch measurement theory (RMT) which goes back to the Danish statistician and measurement pioneer Rasch (1980). Although measurement issues are listed high on the agenda of international marketing researchers, and despite enjoying an increasing popularity in disciplines such as psychology or rehabilitation medicine this particular theory has not yet received the attention in International Marketing it possibly deserves.

In order to examine the usefulness and potential contribution of RMT we examine the construct "technophobia" in a multinational context. We employ both the popular multigroup confirmatory factor analysis approach (MG-CFA) and the Rasch methodology to a dataset of more than 900 respondents from the UK, Mexico and Austria. Technophobia has been introduced by Sinkovics *et al.* (2002) to assess a negative, anxious (phobic) attitude of consumers towards certain innovative products, which results in their being less open to these products, feeling uncomfortable when using them and disregarding technological benefits related to their use. The reluctance of consumers to buy new technology-driven products represents a key threat to a successful and fast diffusion of the market. In the context of international marketing, the awareness of different levels of technophobia can affect market entry decisions and may help design marketing actions to overcome possibly substantial levels of technophobia.

The present study differs from previous research in two significant ways. Firstly, we transcend mostly conceptual discussions regarding advantages of one approach over the other (Ewing *et al.*, 2005; Fan, 1998; Salzberger *et al.*, 1999; Singh, 2004) and address their empirical application. Rather than advocating a particular approach a priori, we attempt to compare approaches empirically. The multigroup CFA approach has been widely used in marketing and international business (Steenkamp and Baumgartner, 1998) and is considered the state-of-the-art methodology in international research. The alternative measurement approach based on Rasch (1980) is still largely underutilized, however, the methodology builds upon a more fundamental definition of measurement. Secondly, rather than using simulated data (Meade and Lautenschlager, 2004; Salzberger *et al.*, 1999), we present a real set of multi-national data. This serves the need of many marketing practitioners and researchers alike to experience the methodology "in real life" rather than in controlled laboratory settings.

2. Problem of data equivalence

International research almost inevitably crosses cultural boundaries. Admittedly, nationality is by no means an ideal definition of culture. However, for the present purpose, nationality serves as an acceptable approximation of cultural affiliation. We do not investigate the sub-cultural level within a nation but we are aware of the fact

that intra-cultural differences might be as relevant as cross-national factors. The point is when multiple populations are to be compared, the issue of data equivalence (Salzberger *et al.*, 1999), also referred to as measurement invariance (Vandenberg, 2002), emerges as a relevant matter of investigation. Data are equivalent across populations when measures bear the same meaning. Therefore, data equivalence is the prerequisite of comparability. Disregarding the question of data equivalence may lead to wrong conclusions. Firstly, observed differences in measures might be attributed to substantial differences between populations although the observed differences are solely caused by differential response behaviour and not by real differences in the latent variables. Secondly, true differences might be masked by differential response behaviour and remain undetected. Consequently, in mean comparisons, both type one and type two errors are increased in an uncontrollable way.

The most prominent case of multiple populations pertains to cross-cultural data. The need to establish and test for cross-cultural equivalence has become accepted in intercultural marketing research. However, there is no final consensus on a specific methodological approach to determine data equivalence. Since, data equivalence refers to differences in quantity as a result of differences in quality, we have to deal with both qualitative and quantitative issues (Salzberger *et al.*, 1999). A typical cross-cultural study entails the application of a scale that is translated into a different language. If the wording of a translated version of an item has a slightly different meaning leading to, say, a higher manifest score compared to the original version given the same latent score, then the qualitative difference in the meaning of the item causes a quantitatively different response depending on the language group. Consequently, every effort has to be taken to ensure comparability of the data during preparation of the instrument (e.g. appropriate translation techniques), administration of the data collection (e.g. comparable setting of interviewing respondents) and so forth. It should be noted, however, that we finally need to empirically verify the actual equivalence of the data.

While there is agreement on the necessity of addressing data equivalence explicitly and quantitatively rather than merely examining the administration and the design of the study in a qualitative fashion, there are different approaches to actually test for data equivalence. There is a good reason why this issue cannot be resolved easily. Dealing with quantitative differences in the measures of different groups can, by definition, not be disentangled from the very issue of measurement itself. Firstly, we have to decide on what theory of measurement we want to rely on, before we can analyse potentially different functioning of items.

3. Conceptual foundations

3.1 Approaches to test for data equivalence

Currently, there are two approaches identified as best practices for assessing equivalence (Reise *et al.*, 1993; Schaffer and Riordan, 2003). Within the paradigm of CTT (Churchill, 1979; Lord and Novick, 1968), the MG-CFA lends itself to testing data equivalence. Introductions into the MG-CFA approach are widely available both within the marketing domain and other disciplines (Cheung and Rensvold, 1998; Steenkamp and Baumgartner, 1998; Vandenberg, 2002). As for the international marketing research community, Steenkamp and Baumgartner, 1998 contribution enjoys a particularly high diffusion and the MG-CFA has become an established approach in marketing research over the recent years.

The alternative approach is based on item response theory (IRT), however IRT applications remain scarce exceptions. The reason is the slow penetration of the measurement theory itself. Embretson and Reise (2000) provide an excellent general introduction to IRT. We, therefore, concentrate on the most important issues and, in particular, focus on those aspects wherein the methods differ substantially.

3.2 Assessing data equivalence by multigroup confirmatory factor analysis

Under CTT, the observed response x_i is a linear combination of a true score and error. In the factor analytic model this translates to an observed response x_i explained by j latent variables (ξ_j) multiplied by the factor loadings λ_{ij}, an intercept τ_i, and an error term δ_I (Meade and Lautenschlager, 2004). Even in the case of multidimensional constructs, we usually relate one item uniquely to one factor (or latent variable). Without considering a person index, in a unidimensional model, the formula then simplifies to:

$$X_i = \tau_i + \lambda_i \xi + \delta_I \tag{1}$$

In one-group studies we usually disregard the intercept term τ_i because it is a constant across respondents and has no impact on (co-)variances. However, if more than one group is considered, differences in item intercept values for the same item in different groups do have an impact on the groups' means. Thus, it is absolutely essential to model item mean vectors and to address this issue in the investigation of data equivalence.

The investigation of data equivalence is carried out by evaluating models that are increasingly stringent (Steenkamp and Baumgartner, 1998). In the baseline model, for all groups considered, the same structure is imposed, i.e. all items are allocated to the same factors and the remaining loadings (non-salient loadings) are fixed to zero. It should be noted that the set of items need not necessarily be the same across groups, even though standard software such as LISREL (Jöreskog and Dag, 2003) require an identical set of items. Baumgartner and Steenkamp (1998) showed that imaginary variables with means of zero, variances of one and covariances of zero with all other variables can be introduced to balance an unequal number of items. The model has to be evaluated following the usual recommendations for judging CFA models. Provided fit is satisfactory, configural invariance is said to hold. Subsequently, constraints are imposed on the loading parameters, specifying the metric invariance model, which establishes a common metric across groups. The decrease in fit can be easily evaluated by a chi-square-difference test with degrees of freedom equal to the difference in the degrees of freedom of the two nested models. While metric invariance may hold for some items, it might not for others. The modification indices point out which items' loadings should be estimated unconstrained. The final model implies partial metric equivalence. For mean comparisons, the origin of the scale of the latent variable has to be defined in the same way for all groups. This requires constraints on the item intercept estimates for those items, which are metrically invariant. This model is termed the scalar invariance model.

Further tests of invariance of error variances and factor (co-)variances can be based on scalar invariance but are not essential for establishing data equivalence. Particularly interesting in the context of intercultural comparisons are constraints on the means of the latent variables. Such a model in comparison with the scalar invariance model can be used to test the hypothesis of equal means across groups.

3.3 Assessing data equivalence based on item response theory

Models based on CTT refer to aggregate statistics like variances, covariances and means. In contrast, in IRT, the manifest response is modelled directly in a probabilistic way. In other words, the model refers to the probability of responding positively, i.e. agreeing, coded as 1, as opposed to disagreeing, coded as 0, in a dichotomous item. In the Rasch (1980) model for dichotomous data (equation (2)), the response depends on person characteristics, covered by the person parameter β_v, as well as item characteristics, operationalised by the item parameter (δ_i). In a general context of measurement, we suggest using the neutral term item location parameter. This parameter has its closest parallel in the intercept parameter in CFA. However, there are fundamental differences. Firstly, the item location parameter is placed on the same scale as the person parameter. Consequently, item and person parameters can be compared directly. Secondly, a reasonable range of item locations in a scale is essential for determining the fit of the data to the model. Moreover, the hierarchy of items helps better understand the construct and adds to construct validity. So, we are always interested in the item location parameters while we usually ignore the intercept parameters in single group factor analysis. The one-parameter logistic model, also known as the Rasch model, is confined to this type of item parameter. We will concentrate on this model because it has some unique features which make it very attractive for measurement in social sciences.

More comprehensive IRT models (see Embretson and Reise, 2000 for an overview) introduce a further item parameter a_i, the item discrimination parameter. However, since varying item discrimination is incompatible with properties, which are important for measurement (primarily specific objectivity), we concentrate on the Rasch (1980) model (equation (2)). In this model, the a_i parameter is dropped from the equation tantamount to a discrimination parameter of 1 for all items.

$$P(X_{vi} = 1) = \frac{e^{(\beta_v - \delta_i)}}{[1 + e^{(\beta_v - \delta_i)}]} \tag{2}$$

x_{vi}, response of person v to item i; β_v, person location parameter; δ_i, item location parameter (endorseability).

For polytomous data, the generalisation of the model is straightforward. Since, there are no assumptions about the scale level, in particular, the response scales are not assumed to be interval scales, polytomous items are characterised by a set of threshold parameters stating the boundaries of adjacent categories. For example, in a five-point rating scale type item, the first threshold tells us where the probability of the first category is equal to the probability of selecting the second. With m categories, we therefore need to estimate a set of $m - 1$ threshold parameters for each item. The threshold parameters can be constrained to be equal across items (rating scale model, Andrich, 1978) or be estimated independently for each item (partial credit model, Masters, 1982). Equation (3) states the formula of the general Rasch model for polytomous data (Andrich, 1988).

$$P(a_{vi} = x|\beta_v, \tau_{ij}, j = 1..m, 0 < x \leq m) = \frac{e^{\left(\sum_{j=1}^{x} - \tau_{ij}\right) + x \cdot (\beta_v - \delta_i)}}{\gamma} \tag{3}$$

with:

$$\gamma = 1 + \sum_{k=1}^{m} e^{\left(\sum_{j=1}^{k} - \tau_{ij}\right) + k \cdot (\beta_v - \delta_i)}$$

a_{vi}, answer of person v to item I (item score); β_v, person v location parameter; δ_i, item I location parameter; τ_{ij}, threshold parameter j of item i; m maximum score, i.e. number of categories

The most important merit of the Rasch model is the feature of specific objectivity (Rasch, 1961, 1977). In essence it says that the item parameter estimates and the person parameter estimates are independent from one another, provided the data fit the model. In other words, the model has to be invariant against all possible groupings of respondents. The invariance property is a defining feature of the class of Rasch models. The relationship between invariance as a model feature and as a property of the data is crucial. According to Fan (1998):

> ... [t]he invariance property (...) makes it theoretically possible to solve some important measurement problems (...) such as (...) test equating and computerized adaptive testing.

However, the more fundamental question is whether measurement has been achieved, at all. Fan (1998) is concerned that the invariance property has been little explored empirically. He seems to suggest that invariance is a property IRT models are said to deliver but that may not hold in reality questioning the value of the model. From a Rasch perspective, the measurement model requires the invariance property of the data in order to provide measures that are comparable, i.e. measures that are on the same scale. Thus, the invariance property is highly important for measurement. If invariance does not hold in the data, the data lack a fundamental property required for measurement.

A test whether the model remains invariant for different cultural groups is therefore only a special case of testing the data model fit. If, however, an item has a different meaning for respondents from different cultures, then the item has a different location or may not even fit at all in one or more groups. Such a non-invariant item is said to display item differential functioning (DIF). A formal test of DIF can be based on the residuals in different groups. In a pooled data set, the mean of the residuals (i.e. the difference between the expected item score and the actual score) is zero. If there is no DIF, this also holds in subgroups, except for random variations. In case of DIF, the residuals deviate systematically from zero, i.e. the mean is positive in one group but negative in another group. A two-way ANOVA lends itself to test the difference for significance. One factor is the group, while the second is the class interval along the latent scale of parameters (Andrich *et al.*, 2003a, b).

In the absence of any DIF, full scalar equivalence holds and data equivalence is given. Like in the CFA approach, the Rasch model also allows for partial equivalence. An item displaying DIF can be split into several versions for each group. That way, a group specific item location parameter is estimated for each item affected by DIF. Moreover, it is very easy to retain an item for one group but discard the item for another. Splitting an item and subsequent deletion of the item in one culture can be carried out conveniently in standard software like RUMM 2020 (Andrich *et al.*, 2003b) whereas the equivalent in structural equation modeling software requires a new set up of the input matrices (Baumgartner and Steenkamp, 1998).

3.4 Contrasting the approaches

3.4.1 Theoretical comparison. Both, the CFA based approach and the Rasch based method to test for data equivalence are appropriate within their respective context. Thus, the evaluation of the competing approaches is best based on the underlying measurement theories. Ewing *et al.* (2005) undertook a comprehensive theoretical comparison and concluded that the Rasch model provides the more powerful basis of measurement in the social sciences. Singh (2004) contrasted CTT and a more general IRT model and pointed out that there are several issues that differ substantially. For example, IRT aims at establishing a broad bandwidth of the instrument, i.e. the scale should include items of widely varying locations. According to Singh (2004) this comes at the expense of reducing fidelity. Singh refrains from favouring either methodological framework but views them as complementary.

Regarding the prerequisites data have to meet, the Rasch model offers some interesting advantages. It can easily be applied to all possible scale formats. Dichotomous data as well as polytomous data or any combination of items with different response formats can all be treated in the same way. In particular, there are no assumptions made about the scale level. In contrast, in CFA we usually do assume a metric scale even though we know that this is extremely doubtful. A further advantage of the Rasch model is the independence of the distribution of respondents, which need not be normal or meet any other predefined shape. The stringency of the Rasch model referring to item discrimination favours CFA, which allows for item specific discrimination. It should be noted that more complex IRT models do incorporate discrimination parameters. In the Rasch model, discrimination is constrained to be equal across items for theoretical and philosophical reasons (Ewing *et al.*, 2005).

An evident benefit of the MG-CFA is its embedding in the standard measurement theory in marketing research - irrespective of possible theoretical virtues of the Rasch approach. If one wants to remain within the paradigm of CTT, the MG-CFA approach is appropriate and it will certainly prosper in marketing research. No clear differences can be found for sample sizes. In contrast to more complicated IRT models, the Rasch model works with about the minimum number of respondents usually recommended for CFA studies. The software to estimate the models has become more user friendly in both cases, so we cannot see any reason to favour either method simply because of the user-friendliness of the software.

Another issue is the handling of items displaying DIF. Accounting for partial invariance by allowing the item to vary across cultures can be done easily in both cases. Sometimes, however, an item may fit on one culture but misfit in another. In a MG-CFA study such an item is typically discarded even though Baumgartner and Steenkamp, 1998 show a way to overcome this problem. The Rasch model offers a different resort. Since, the input to the Rasch model is the raw data rather than aggregated statistics like covariance matrices, missing data represents no substantial problem. While missing data do increase standard errors of parameter estimates, the estimation is not affected, in principle. That is why an item parameter may easily be estimated for one group while the same item is discarded for other groups. Consequently, the inclusion of culture-specific items (emics) is easier with the Rasch model, both from a conceptual and operational point of view.

3.4.2 Empirical comparison. Only the application to empirical data can shed light on how the theoretical differences bear on the conclusions drawn from analyses of data

equivalence. Ewing *et al.* (2005) demonstrate how the Rasch model can be applied to a multi-group set of real data. However, no comparative analysis with the traditional CFA approach has been undertaken. Nevertheless, the study shows that Rasch analysis is powerful in analysing typical marketing research data.

Salzberger *et al.* (1999) compared Rasch analysis and the MG-CFA approach using a simulated data set. The study, which is mainly a conceptual contribution, illustrates how the different models work in a situation where one subset of items is affected by the same additive bias in one group while another subset of items is invariant across groups. Since, bias is always relative, either subset of items may be considered to be invariant with the other subset being biased. In this case, the MG-CFA approach is very unlikely to recover this fact while the Rasch analysis typically reveals this fact. The reason is that the MG-CFA approach relies on one particular item for definition of the latent scale while the Rasch model defines the scale origin as the mean of all item locations. An important conclusion from this study is the fact that the statistical analysis of data equivalence should always be complemented by substantial content-related considerations.

A very comprehensive comparative study is provided by Meade and Lautenschlager (2004). The authors compare the MG-CFA method with an IRT model that allows for different item discrimination, i.e. a non-Rasch model. Based on a series of simulated data sets varying sample sizes and the type of DIF (location parameter DIF and item slope DIF, respectively), Meade and Lautenschlager (2004) conclude that the IRT approach is "somewhat better suited for detecting differences when they were known to exist". They also emphasize that "CFA methods … were inadequate at detecting items with differences in only *b* parameters." and "CFA methods were also largely inadequate at detecting differences in item a parameters." This seems to be a strong case in favour of the IRT approach. However, as the authors concede in their discussion of limitations, the way the data is simulated is crucial. Meade and Lautenschlager (2004) used IRT-based software (Baker, 1994). This implies that the responses are reflecting a non-linear relationship of the latent variable and the manifest responses. Consequently, a basic assumption of the CFA approach, namely that there is a linear relationship, is violated suggesting the CFA approach is inappropriate in the first place. Still, the conclusions of Meade and Lautenschlager (2004) should be considered valid. The reason is that real data, if they are suitable for measurement, should be non-linear because responses are bounded between a limited number of response categories and item locations (or item intercepts) are varying between items. Thus, Meade and Lautenschlager (2004) illustrate how the CFA model behaves with reasonable data.

The issue of data generation also indicates that there actually can be no "theory-free" comparison of approaches which differ substantially in their foundation. One can either draw conclusions purely on a theoretical basis or one can refer to data. If the data are real data, one does not know which items, if any, are affected by DIF. Then, different results cannot be interpreted without reference to theoretical considerations. If the data are simulated, one has to decide according to which model the data are generated. So, the theoretical input creeps into the study at this stage. However, data are generated, one has to argue for the model chosen. Thereby, one postulates how data should look like and how they should be structured. However, this claim is tantamount to favouring one model over the other, what in turn is a theoretical decision.

4. The empirical study

4.1 Purpose of the comparison

The present study refers to a comparison of a Rasch based approach and the MG-CFA approach to test for data equivalence in a set of real data. Naturally, practical marketing research is concerned with real data. We never know if data are equivalent or where DIF occurs. We even do not know a priori that measurement has been achieved at all. Data may contain so much error that we have to reject the notion of quantification. Although rarely considered possible, a variable may not even exist in quantity, then measurement ceases to be meaningful altogether.

Simulation studies like those conducted by Meade and Lautenschlager (2004) or Salzberger *et al.* (1999) as well as applications of Rasch models to cross-cultural data (Ewing *et al.*, 2005) and intra-cultural data demonstrate that Rasch and IRT methods do represent an interesting option of conceptualising measurement. Based on theoretical considerations, we frame the following expectations. We refrain from calling them hypotheses simply because there is no unambiguous theoretical argument but often pros and cons.

Expectation E1.
In general, we expect more items to fit the CFA model compared to the Rasch model.

It is often argued that the Rasch model is more demanding than the traditional test theory. In particular, items are required to be equally discriminating under the Rasch model. This should lead to more misfits when analysing data using the Rasch model. In particular, items deviating strongly from the mean discrimination (factor loading), should misfit the Rasch model. However, if the items vary substantially in their location, then floor and ceiling effects can lead to reduced item-covariances. Then the items may even fit better under the Rasch model. Since, the items have not been generated with an eye on maximising item locations, we do not expect this effect to play a substantial role, though.

Expectation E2.
In the analysis of data equivalence, we expect a similarity of items lacking scalar invariance in the MG-CFA approach and displaying DIF in the Rasch approach.

Item intercepts and the item locations are related parameters. If items are non-invariant, item intercepts should differ between groups and items should display DIF. In general, the item intercept parameters and the item locations should be inversely correlated. The harder an item is agreed with, the larger the item location and the smaller the item intercept because the manifest score is smaller compared to an easier item given the same latent score.

Expectation E3.
For items lacking fit across groups, we expect some to fit in at least one group but not in another. Therefore, under the Rasch model, the number of items in the final scale should be larger, all other things being equal.

This expectation differs from the two before. In the Rasch model it is very easy to retain an item for one group but discard it for another. The question is whether there are such items in the scale, which is a substantial problem rather than a methodological

one. Nevertheless, the inclusion of items unique to one group may balance the effect of expectation E1.

4.2 Conceptual foundation of technophobia

The Anglo-American literature offers a multitude of conceptual foundations for technophobia, particularly pertaining to synonyms such as techno stress (Brod, 1984), cyberphobia (Price and Ridgeway, 1983), computer aversion (Meier, 1985) or computer anxiety (Raub, 1982). The findings however, are largely restricted to "computer"-phobia, as computers were used as anchor products. Scholars have argued that computerphobia and technophobia relate to the same latent variable (Rosen and Weil 1990a, b). However, in view of potential generalization problems in the international context, Sinkovics *et al.*, 2002 established a generic technophobia scale that is applicable to a variety of products and services. The scale is exemplified by referring to automated teller machines.

Their instrument is deemed to represent negative psychological reactions towards technology, which can arise in various forms and intensity. Hereby the term 'phobia' is not used in a strict medical sense, relating to the results of the exposure to a feared situation (often demonstrated in symptoms such as sweating, tremors, flushing, etc.), but the notion of phobia implies rational (Röglin, 1994) and – what is even more – irrational psychological aspects to the anxiety (Jaufmann, 1991). The authors derive a three-dimensional factor-structure for technophobia (Sinkovics *et al.*, 2002). The first factor relates to "personal failure", i.e. issues describing problems, frustrations, and failures when using sophisticated or innovative machines, the second factor represents issues which elicit the ambiguity between human and machine interaction, i.e. fears about machines dominating interactions. Lastly, the third factor is related to "convenience" issues, when using machines. The three factors correlate between 0.45 and 0.59.

4.3 Data

To illustrate the procedures for equivalence testing, a subset of data originally collected for a large multi-country survey of consumers was used (Sinkovics *et al.*, 2002). The original study established a measure for the concept of "technophobia" and comprised additional measures such as "innovativeness" (Hirschman, 1980; Price and Ridgeway, 1983). The items used five-category Likert scales. Sinkovics *et al.*, 2002 developed the "technophobia" scale and found reasonably well reliability scores and indications for validity, following exploratory and confirmatory multi-group structural equations modelling procedures. In terms of controlling for equivalence (Craig and Douglas, 2005), different sampling frames were employed. In Great Britain, the sample was drawn from four metropolitan areas, in Mexico, a student sample was taken and in Austria a quota sample was drawn which was representative for the adult Austrian population. Quota descriptors included age, gender and occupation.

4.4 Descriptive results

Our analysis builds on data (total $n = 927$) from the UK ($n = 278$), Mexico ($n = 200$) and Austria ($n = 449$) in a deliberate attempt to maximize some of the outcomes of the methodological comparison. Sample sizes were equally high in all three countries, while at the same time we were dealing with fundamentally different cultural

environments. English, Spanish, and German languages were represented in the data and different perspectives of the technophobia phenomenon were expected.

Given the sampling procedures described above, there existed a slight bias in terms of younger age group ($F = 8.577$, df $= 2$, $p < 0.001$). Another potential bias was introduced by the fact that data collection was mainly administered in urban regions. However, these effects were not considered to threaten the methodological direction of our research, quite contrastingly, since these were structural biases of the samples, consistent across countries, the comparison between CFA and Rasch was deemed to be even more interesting.

While card ownership is generally wide-spread in the countries under scrutiny, it is significantly higher in the UK-sample (93.1 percent have ATM cards) than in the Austrian and Mexican sample with 75.9 and 73.0 percent, respectively. The same pattern prevails for the frequency of card usage. The differences are considered to be the result of unequal stages of technological development and product diffusion patterns for automated teller machines.

4.5 Results of the MG-CFA analysis

The MG-CFA was conducted following the scheme recommended by Steenkamp and Baumgartner (1998) using Lisrel 8.54 (Jöreskog and Dag, 2003). The starting point was a multi-group model without any constraints across groups other than the mere factorial structure, i.e. a unidimensional model. We attempted to keep the scale undimensional for the sake of parsimony. However, if the data had indicated that this notion is not viable, we would have switched to a multidimensional model.

In assessing fit, we mainly focussed on the RMSEA and the χ^2-statistic. The first model included all 30 items. The fit of the model was highly unsatisfactory. Subsequently, the modification indices were used to identify items the error terms of which show significant covariation. In most cases these items were similar in content. Consequently, one of them was deleted. This procedure was carried out iteratively until a model was derived that fitted well. Finally, a set of six items turned out to fit in all three groups under scrutiny. This model established configural invariance. It acts as the baseline model against the more stringent model of metric invariance is tested.

The next step consisted in imposing equality constraints on the loading parameters across groups. The resulting model formalises full metric equivalence across groups ensuring the same unit of measurement prevails in all three countries. The difference in the χ^2-statistics of the configural and the metric invariance model can be evaluated by the χ^2-difference test with degrees of freedom equal to the difference in degrees of freedom of both models. The decrease in fit of the metric invariance model turned out to be non-significant implying full metric invariance to hold true in the three samples considered. All tests were evaluated on the 1 percent-level for type one error.

Subsequently, the item intercept estimates were constrained to be equal across groups in the full scalar invariance model. These further constraints resulted in a significant decrease in model fit compared to the metric invariance model indicating that some items are affected by non-invariance. Again, the modification indices of the intercept parameters showed us where we had to lift constraints. For one item a unique item intercept had to be estimated in each group while another item required a separate estimate in the Austrian sample but worked well with a common estimate for the UK and Mexico.

In summary, the MG-CFA approach established a relatively high degree of invariance given the diverse nature of the countries. Metric invariance prevails fully four items are scalar invariant. It should be noted, though, that 24 items were discarded and only six items retained. A set of six items should be sufficiently precise for most applications and it is certainly highly economical concerning the expenses during data collection. However, the question arises whether six items out of 30 adequately represent the construct. When looking at the content of those items that were retained and of those that were discarded, it is striking that the retained items are primarily about the ease of using ATMs (worry about making mistakes, easy to learn, time-consuming, etc.) while only one item (I do not trust ATMs with my money.) is emotionally coloured.

Based on the model of partial scalar invariance, further constraints can be imposed, which are not necessary for data equivalence in terms of mean comparability but provide interesting information. First of all, a latent mean test can be conducted by requiring the latent group means of all three samples to be equal. While the latent means differed significantly between all three groups, a subsequent test of equal latent means of the UK and the Mexican sample was insignificant suggesting that technophobia is significantly smaller in Austria versus the UK and Mexico with no significant difference between the latter countries.

Invariance was further extended by forcing item loadings to be equal across items in addition to equal loadings across groups. This is tantamount to equal item discrimination across items. A model of partial scalar equivalence with equal item discrimination for three items could be established. Alternatively, error variances were selected to be equal across groups. Some equality constraints of this type had to be lifted but finally a partial scalar invariance model with equal error variances fitted the data well. These further tests of invariance demonstrate that a high degree of equivalence prevails in the data clearly exceeding the absolutely necessary level of comparability (Table I).

4.6 Results of the Rasch analysis

The Rasch analysis is based on the partial credit model (Andrich, 1988; Masters, 1982), i.e. each item is assumed to have its own set of distances between category thresholds. The additional advantage of the Rasch approach is that the threshold ordering of these items can be empirically examined. Although constructed to imply an increasing level of the latent dimension, in practice the response categories may not work properly and the threshold estimates may take any order suggesting violations of the hypothesis that the response data are ordinal. In the present study threshold disordering did occur. A *post hoc* remedy to cope with this issue is collapsing adjacent categories. For this reason, in the current data set, all items were rescored by retaining the first category (fully disagree), collapsing the second category (disagree) and the middle category, and by collapsing the remaining categories of agree and fully agree.

The assessment of data equivalence using the Rasch model features some similarities but also exhibits some differences. Since, all items are required to discriminate equally, metric equivalence can only be achieved fully. Scalar invariance, however, can be partial insofar as an item may display differential item functioning (DIF) requiring a split of the item between the groups. An important difference lies in the fact that the Rasch analysis starts with a full invariance model. Consequently,

Type of invariance	Model/constraints	RMSEA	χ^2 (df)	p value	$\Delta\chi^2$ (df)	p value
Starting model	No constraints across groups	0.115	4057.30 (1131)	<0.0001	–	–
Deletion of items due to significant covariation of error terms						
Configural invariance	No constraints across groups	0.046	43.11 (27)	0.025	–	–
Full metric invariance	Constraints on loadings (λ_g)	0.043	58.22 (37)	0.015	15.11 (10)	0.128
Full scalar invariance	Constraints on item intercepts (τ_{jg})	0.093	154.16 (47)	<0.0001	95.94 (10)	<0.0001
Setting free item intercept parameters						
Partial scalar invariance	Constraints on some item intercepts (τ_{jg})	0.046	71.35 (44)	0.006	13.13 (7)	0.069
Latent mean test	Constraints on all latent means	0.068	107.87 (46)	<0.0001	36.52 (2)	<0.0001
Latent mean test Austria versus (the UK and Mexcio)	Constraints on latent means in the UK and Mexico	0.046	73.25 (45)	0.005	1.9 (1)	0.168
Full metric invariance plus equal discrimination of three items	Additional constraints on loadings across three items (λ_j)	0.041	58.88 (39)	0.021	0.66 (2)	0.719
Partial scalar invariance plus equal discrimination of three items	Additional constraints on loadings across three items (λ_j)	0.041	72.19 (46)	0.008	0.84 (2)	0.657
Partial scalar invariance plus some equal error variances	Additional constraints on error variances (for two items for all three countries; for two items for Austria and the UK; for two items for Austria and Mexico)	0.041	89.09 (52)	0.001	17.73 (8)	0.0233

Table I.
Assessment of data equivalence with the MG-CFA approach

the item parameters are estimated from pooled data. The overall fit of the model can be assessed by a χ^2-test that compares, for each item, the expected scores based on probabilities with actual scores based on proportions in several score groups, added up over all items (Andrich *et al.*, 2003a). The test can also be used for the evaluation of an individual item. Basically, misfit of an item can have two reasons. Either the item misfits in at least one of the groups or the item exhibits DIF. Therefore, at each step, item fit statistics have to be examined carefully as well as the test for DIF. Besides, the person separation index was monitored. This statistic is similar to classical reliability. It is bounded between zero and one. High values mean that the items discriminate between the persons. Small values (below 0.85) are problematic because the test of fit loses power. All analyses were conducted using RUMM 2020 (Andrich *et al.*, 2003b).

The analysis was carried out iteratively. At each step, only one change to the model was undertaken in order to disentangle the effects of misfit of different items. Either an item was split up because of DIF or an item was discarded (for all countries or, after a split up, only for one country). The final model comprised a total of 13 items. However, only four of these were invariant across all three groups. Four other items were invariant across two groups with one case were the item misfitted in the remaining group. Two items had to be split up for each group due to DIF. Two items fitted only in two groups and exhibited DIF, one item fitted only in Austria. In summary, 12 items are available for Austrian respondents, whereas the scale comprises 11 items for respondents from the UK and Mexico, respectively.

Since, the test of DIF based on the analysis of variance may suffer from the unequal sample sizes, the results were scrutinized carefully. Firstly, unequal sample sizes imply different power of detecting DIF involving a particular country. Secondly, the sum of squares cannot be allocated unequivocally to the main effects and the interaction term when the design is not orthogonal. To address the first issue, we screened the DIF findings. In principle, the power of the test of DIF is expected to be higher when Austria is involved because it has the largest sample size. However, this does not result in more items showing DIF for Austria in our example. There are three items that are interesting in this respect because they fit in all three countries. In one case, there is a common estimate for the UK and Mexico (in line with higher power for differences against Austria), in another case there is a common estimate for Austria and Mexico, and finally, in the third case there is a common estimate for Austria and the UK. So, it seems that the issue of unequal power does not bear on the outcome in the Rasch analysis. In the traditional CFA analysis, in one case, there is a separate estimate for Austria and a common one for Mexico and the UK.

To examine the second potential threat to the DIF analyses, a random subsample of the data set was drawn with 194 respondents from Mexico and 200 from Austria and the UK, respectively. Since, the DIF diagnosis did not depart in any way from the findings based on the complete data set, we consider the findings tenable. Apart from the DIF analyses, the sample size also has an impact on fit statistics. When assessing the overall fit in Rasch models, the sample size and the person separation index have to be considered particularly. In the present data set, 817 respondents are available for fit assessment. The remaining respondents display extreme scores and, therefore, do not qualify for fit analyses. In the context of Rasch models, 817 represent a large number. Moreover, the person separation index amounts to 0.89 implying excellent power of the test of fit. This leads to a very sensitive χ^2-statistic. We, therefore, focussed on

the incremental improvement of fit when deleting or modifying items at each step. Another indication of fit was the fit statistics for each individual item. Even the worst fitting item exceeded the 1 percent level. Furthermore, RUMM provides the option of calculating adjusted χ^2-statistics based on a different sample size. Boomsma and Hoogland (2001) mention $n = 200$ as the minimum sample size when discussing robustness of structural equation modeling against small sample sizes. If we consider $n = 200$ as the minimum sample per country, which is also a reasonable minimum sample size in a Rasch analysis, we get a total sample size of 600. With 600 respondents the χ^2-statistic amounts to 113.07 (df $= 88$, $p = 0.037$). However, greater confidence can be derived by taking random samples of 600 respondents and repeatedly calculating fit statistics like in a bootstrapping approach. With a mean χ^2 of 126.35 (df $= 88$, $p = 0.005$) from 30 runs the model fit appears to be marginal. Analysing each country separately, the data fit the model satisfactorily (Austria $\chi^2 = 72.63$, df $= 48$, $p = 0.012$; UK $\chi^2 = 66.90$, df $= 44$, $p = 0.015$; Mexico $\chi^2 = 53.90$, df $= 44$, $p = 0.145$). For these reasons we deem the final model tenable (Table II).

Figure 1 shows the main results of the Rasch analysis. It depicts the person locations (displayed on top) against the item threshold locations (displayed downwards), which are placed on the same dimension. The difference between Austria and the UK and Mexico, respectively, is about one logit unit and statistically significant. The bell-shaped curve shows the amount of information the instrument provides for respondents depending on the level of technophobia. The more information we have about a person, the smaller is the standard error of the person location tantamount to higher measurement precision. The distribution of the respondents is shifted to the left, implying that for many persons the items are relatively hard to agree with. Consequently, the scale is more sensitive in the area of moderate to severe technophobia. However, that is exactly what the scale is supposed to be. We do not want to differentiate between respondents of negligible degrees of technophobia.

4.7 Comparison of the results

The comparison of the results gained from either approach shows some striking parallels (see Table III). Each approach suggests four items, which are fully equivalent. Three of these are reported invariant (i.e. v07, v27, and v29) in both analyses. Another item (v18) displays a relatively small but significant additive bias in the Austrian sample against the other two samples whereas the Rasch analysis could not find any indication of DIF (the ANOVA DIF test is insignificant with $p = 0.32$). Item 14 is particularly interesting since it turns out to be fully invariant in the MG-CFA approach. In contrast, in the Rasch model it fails to meet a reasonable level of fit and also shows strong indication of DIF. The item is special insofar as it is one of the few reverse coded items. Such items are sometimes problematic since they are prone to confuse respondents. In many cases people do not simply respond in a "reverse" way. Instead, they favour extreme categories and make less use of the finer distinctions in between. In the Rasch model this may lead – quite reasonably – to misfit while in the CFA approach the more "pronounced" responses may even enhance the fit of the item. This may have occurred in the present study. Another point is whether the perceived easiness of learning how to use an ATM is actually a good indicator of technophobia.

Type of invariance	Model	Fit statistics/ANOVA statistics χ^2 (df)	*p*value	Fit statistics of the worst fitting item	Person separation index
Full scalar invariance	Common parameter estimation across all groups	2,107.93 (270)	< 0.000001		0.93
Deleting items or splitting up items					
Partial scalar invariance	Some items split up because of DIF	153.96 (88)	0.000018	$\chi^2 = 11.52$ (df = 4) v	0.89
Latent mean test	Analysis of variance of latent means	$F = 18.73$ (6,912)	< 0.000001		
Latent mean test Austria versus (the UK and Mexico)	Analysis of variance of latent means, *post hoc* test	Scheffé test: Austria − 1.584, the UK − 0.505, Mexico − 0.381		Austria vs the UK $p < 0.001$ Austria vs Mexico $p < 0.001$ the UK vs Mexico$p < 0.717$	

Table II.
Assessment of data equivalence with the Rasch partial credit model

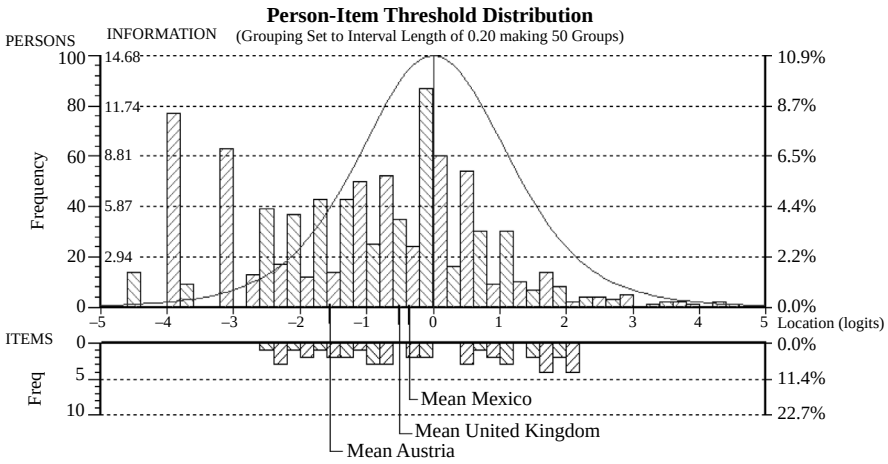

Figure 1.
Rasch Person measures and item threshold parameters

Code	Item wording	MG-CFA (item intercepts) [a] Austria	The UK	Mexico	Rasch analysis (item locations)[b] Austria	The UK	Mexico
v07	Using ATMs is time-consuming	1.93	1.93	1.93	0.393	0.393	0.393
v18	I dont trust ATMs with my money	2.13	2.28	2.28	−0.215	−0.215	−0.215
v27	I find ATMs instructions confusing	2.13	2.13	2.13	0.01	0.01	0.01
v29	I feel confident that I could teach someone how to use an ATM (reverse coding)	3.74	3.74	3.74	−0.267	−0.267	−0.267
v21	I wish I were more adept at using ATMs				0.289	−0.437	
v26	Machines should not handle people's money transactions					−0.203	−0.203
v13	I refuse to use ATMs				−0.196		0.425
v01	I feel some anxiety when I approach an ATM				0.736	−0.091	−0.411
v06	I worry about making mistakes when using ATMs	2.00	2.53	2.91	−0.067	−0.864	−1.273
v17	Thinking about ATMs makes me nervous				1.347		
v09	It takes me a long time to complete bank transactions when using an ATM				0.684	0.079	0.079
v08	ATMs agitate me				0.528	−0.111	0.528
v23	ATMs seem very complicated				0.224	0.224	−0.582
v14	It is easy to learn how to use ATMs (reverse coding)	3.93	3.93	3.93			
	Reliability (Cronbach's α)	0.76	0.83	0.72	0.875	0.884	0.879

Table III.
Results of assessing data equivalence

Notes: [a]The parameter estimates from the partial scalar invariance model are shown here without any further equality constraints on loadings across items or error variances; [b]In the table cells the Rasch overall item locations are stated. These locations are the mean of the two thresholds that are estimated for each item. The mean of all item thresholds over all items is zero by definition

In fact, people might be intelligent and clever and still reject the idea of "machines handling money transactions".

By both approaches, item v06 is reported to be an indicator of technophobia in all three countries requiring additive correction for DIF. According to the MG-CFA and the Rasch approach, the item is easiest to agree with in Mexico, followed by the UK and finally Austria, representing another parallel between the approaches.

Interestingly, the Rasch methodology reveals eight other items that are at least in one of the countries valid indicators. Four of these items are equivalent, i.e. free of DIF, for two groups providing a stronger link between the countries and enhancing comparability.

Looking at the item intercept estimates in the MG-CFA and the item locations from the Rasch analysis, the same rank order appears for the almost completely invariant items retained under both models. Item v29 is the easiest item, followed by v18, v27, and v07. The metric correlation of all intercept estimates and item locations from common results amounts to -0.47 ($p < 0.01$).

Regarding the comparison of the means of all respondents between the countries, the two approaches come to the same conclusion. Respondents from Austria are less technophobic than people from the UK or Mexico. On the individual level, the correlation of the Rasch person measures and the latent variable estimates from the CFA is 0.85 across all countries. Within the three different groups the correlation coefficient are even higher with $r = 0.89$ in case of Austria and Mexico, and $r = 0.87$ for respondents from the UK. This seems to suggest that Rasch measures can be used interchangeably with CFA derived scores. However, such a conclusion would be premature for at least three reasons. Firstly, Ewing *et al.* (2005) pointed out that the traditional method based on CFA might be an approximation if, and only if the data fit a Rasch model. The Rasch model provides evidence about the data justifying measurement that lies outside the potential of CFA. Secondly, even a correlation as high as 0.89 means that almost 21 percent of the variance is not shared. This issue will be addressed below when we investigate further evidence of validity based on correlations. Thirdly, an important difference between the approaches considered is the non-linearity of the relationship between the measure and the raw score in the Rasch model and the linear relationship in case of CFA. Since, the raw score is transformed in a non-linear way in the Rasch model but in a linear way in CFA, a non-linear relationship prevails between the Rasch measure and the CFA based score. In particular, the non-linearity becomes the most prominent at the extremes. Incidentally, this implies that it is theoretically impossible that both approaches simultaneously yield linear, interval scaled measures. In the present study, only few respondents possess a high degree of technophobia. Consequently, the non-linearity at the upper end plays no substantial role. At the lower end the non-linearity is clearly visible (see Figure 2 for the relationship in the Austrian sample, the figures for the UK and Mexico look very similar). However, the majority of the respondents are located in the central area where the relationship is virtually linear.

4.8 Testing the expectations

In the theoretical section, we raised three expectations (E1-E3) regarding potential differences in the outcome of the analyses. E1 claimed that more items would fit

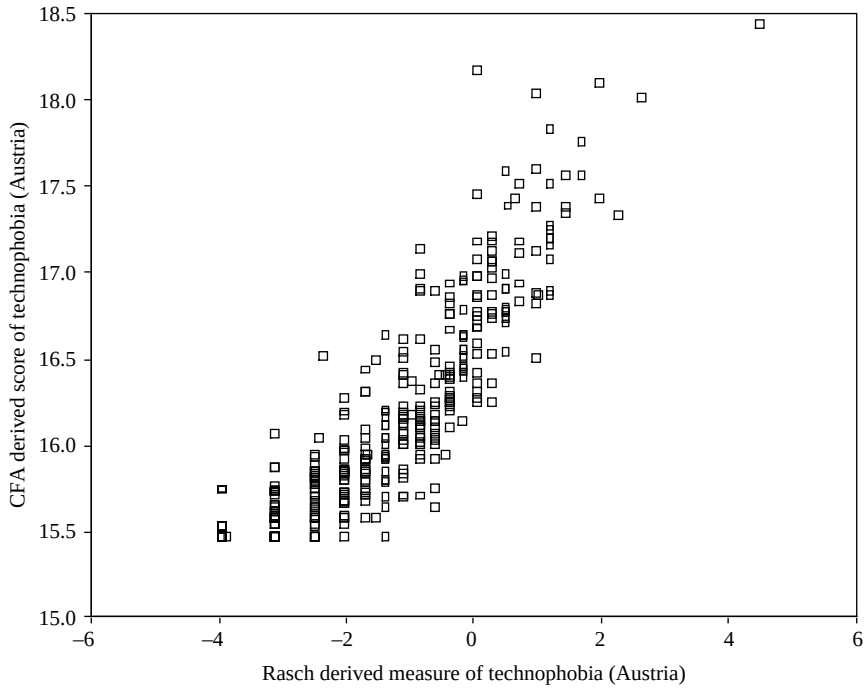

Figure 2.
Relationship of Rasch derived measures and CFA based scores of technophobia

the CFA model compared to the Rasch model. This expectation has to be rejected. In fact, more items, if only partially (see also E3), are retained in the Rasch analysis. While the traditional solution based on CFA has to get by with six items, the Rasch derived scales comprise 11 (the UK and Mexico) and 12 items (Austria), respectively. In turn, the Rasch model rejects one rather doubtful item, which features even scalar equivalence in the MG-CFA. In total, the Rasch model seems to provide a more justified and precise person measure compared to the CFA even though it does not make a difference on the aggregate level of group means. In terms of item content, the set of items retained in the Rasch model covers the emotional component (anxiety, agitation, nervousness, intimidation) much more comprehensively than the CFA based scale. The reason why the Rasch model fits even better and why, all in all, both approaches display significant parallels are twofold. Firstly, the items show a high degree of equivalence and allow for additional constraints on the loadings across items as well as on the error variances. This is not so much a cross-cultural issue but a general feature of the data that explains why the data fit the Rasch model reasonably well. Secondly, the items have been generated following the classical paradigm that favours high inter-item-correlations. In particular, items were not generated aiming at substantial differences in item intercepts (this is referred to as 'bandwidth' issue by Singh, 2004). In contrast, the Rasch (and also the IRT) philosophy requires sufficient variation in item locations for establishing construct validity (Ewing *et al.*, 2005; Singh, 2004) and

providing enough information across a wide range of the scale. The technophobia scale is certainly at the lower end of the acceptable range of item locations. This also favours the CFA (see Salzberger *et al.*, 1999 for reasons why differences in item intercepts may also have an impact on loadings).

The second expectation E2 stated that there would be a similarity of items lacking scalar invariance in CFA and displaying DIF in Rasch analysis. The results support E2 in principle. Firstly, the reverse is true, i.e. those items that meet scalar invariance requirements do not exhibit DIF. Secondly, one item (v06) shows the same pattern of relative biases in both analyses.

The third expectation, according to which some items lacking fit should fit the Rasch model in some of the groups. E3 is strongly supported. Eight items are part of the Rasch derived instrument but are discarded during the MG-CFA.

4.9 Further evidence of validity

Notwithstanding the somewhat different understanding of construct validity depending on the theoretical foundation of measurement, the CFA as well as the Rasch analysis support the construct validity of the technophobia scale as well as its cross-cultural validity as far as the three countries under scrutiny are concerned. Criterion-based validity provides further evidence of validity. Consequently, the technophobia scores from CFA and from the Rasch analysis were correlated with the "use innovativeness" measure, as established by Price and Ridgeway (Hirschman, 1980; Price and Ridgeway, 1983). The construct deals with the use of previously adopted products in novel ways and encompasses five factors: creativity/curiosity, risk preferences, voluntary simplicity, creative reuse, and multiple use potential. The first four were used in the present study. An index across these dimensions was also calculated and named use innovativeness. The scales were analysed exclusively based on CFA.

Considering the technophobia measure based on MG-CFA within the country groups, nine correlations are significant (Table IV). There are some noteworthy differences between the countries. For example, risk preferences are not significantly correlated with technophobia in the UK but correlate -0.40 in Mexico.

Since, the Rasch derived measure and the CFA based score are highly correlated, it is not surprising that very similar patterns of correlations occur if the Rasch measure (Table V) is used instead of the CFA score. Looking at the nine correlations that are significant in both cases, seven are higher with the Rasch score, one is equally high and only one correlation is lower when the Rasch measure is used. In turn, three correlations are non-significant in both cases, one is lower in case of Rasch, two are lower when the CFA is employed. Finally, one correlation is non-significant with CFA but significant with Rasch. In summary, eight correlations are closer to 1 (significant correlations) or closer to 0 (non-significant correlations) with Rasch. Only three correlations are closer to 1 (significant correlations) or closer to 0 (non-significant correlations) with CFA. The findings cannot be generalised but it seems that the Rasch measures have got higher precision and accuracy sharpening the correlations accordingly.

In any case, the findings confirm the assumption that higher levels of innovativeness go with lesser levels of technophobia.

Pearson correlation with technophobia score from MG-CFA analysis	Creativity, curiosity		Risk preferences		Voluntary simplicity		Creative reuse		Use innovativeness	
	$r(n)$	p value	$r(n)$	p value	$r(n)$	p value	$r(n)$	p value	$r(n)$	p value
The United Kingdom	0.04 (262)	0.58	−0.10 (263)	0.10	−0.29** (266)	<0.01	−0.25** (265)	<0.01	−0.19** (255)	<0.01
Mexico	0.05 (168)	0.52	−0.40** (175)	<0.01	−0.28** (176)	<0.01	−0.24** (177)	<0.01	−0.19* (166)	0.02
Austria	0.08 (438)	0.12	−0.12** (438)	0.01	−0.07 (438)	0.12	−0.10* (438)	0.05	−0.04 (438)	0.40
Pooled data	0.08* (868)	0.02	−0.09** (876)	0.01	−0.25** (880)	<0.01	−0.18** (880)	<0.01	−0.10** (859)	<0.01

Notes: *, **Correlation is significant at the 0.05, 0.01 level (two-tailed), respectively

Table IV.
Correlations "use innovativeness" and "technophobia" based on MG-CFA

Pearson correlation with technophobia score from MG-CFA analysis	Creativity, curiosity		Risk preferences		Voluntary simplicity		Creative reuse		Use innovativeness	
	$r(n)$	p value	$r(n)$	p value	$r(n)$	p value	$r(n)$	p value	$r(n)$	p value
The United Kingdom	−0.01 (269)	0.92	−0.09 (270)	0.14	−0.37** (273)	<0.01	−0.28** (272)	<0.01	−0.26** (262)	<0.01
Mexico	−0.06 (179)	0.40	−0.42** (189)	<0.01	−0.29** (190)	<0.01	−0.24** (192)	<0.01	−0.26** (166)	<0.01
Austria	0.12* (449)	0.02	−0.13** (449)	0.01	−0.08 (449)	0.10	−0.10* (449)	0.05	−0.02 (449)	0.72
Pooled data	0.07* (897)	0.03	−0.17** (908)	<0.01	−0.24** (912)	<0.01	−0.19** (913)	<0.01	−0.13** (887)	<0.01

Notes: *, **Correlation is significant at the 0.05, 0.01 level (two-tailed), respectively

Table V.
Correlations "Use innovativeness" and "technophobia" based on Rasch analysis

5. Summary and implications

The Rasch model and the measurement theory the model is based on, is a promising alternative to the standard approach to measurement rooted in the classical measurement theory. As Ewing *et al.* (2005) have pointed out, RMT can be seen as a more comprehensive framework compared to the classical factor analytic approach. The Rasch model justifies the computation of the raw score that is calculated in either approach in the same way (except for different weighting of the item scores). While the classical theory regards the scores as linear measures, the Rasch model acknowledges the non-linearity and transforms the raw scores into a linear, interval-scaled measure by a logistic function. From this it follows that the Rasch measures and the CFA measures cannot concurrently be linear.

Where the data do not fit the Rasch model, the theoretical foundation of measurement, as it is naturally and routinely asked for in modern physical science, is missing. Resorting to CTT or to more complicated IRT models may or may not lead to a solution. However, we should realise that this sort of measurement is not compatible with the notion of quantification held in the physical sciences. One might argue measurement in the social sciences is harder to achieve. Indeed, this seems to be the case but does it really exempt us from being rigorous and allow us to be more speculative? In fact, one would hardly object to the proposition that numerals without quantitative meaning do not qualify for any sort of statistical computations. Hence, science commits us to provide empirical evidence that numerals reflect quantity.

Where the data do fit the Rasch model, we can be confident that the underlying variable is quantitative and that measurement has been achieved. Of course, this applies within the limits of unavoidable statistical error affecting any sort of empirical hypothesis testing. However, even in this case, the level of manifest item scores is only ordinal. Consequently, the raw data do not meet the assumptions for factor analytic procedures, at least not for those based on covariances or Pearson correlations. On the other hand, the vast majority of empirical research in marketing is based on CTT. It would be overdone to claim that all these findings are invalid. Despite the fundamental philosophical differences between CTT and Rasch measurement, CTT can be seen as an approximation to Rasch measurement – provided the data fit the Rasch model. This implies that many findings would remain valid. What we would "lose" – or rather identify as being fallacious – are those findings where our asserted measures do not reflect quantity. Consequently, when applying the Rasch model there is nothing to lose but much to be won.

In more practical terms, the Rasch model offers further interesting advantages. It can be applied to any number of response categories with all possible combinations within one instrument. The manifest responses are assumed to be ordinal and need not be interval-scaled. The distribution of the respondents can take any shape without endangering parameter estimation, in principle. A fundamental difference between the paradigms of RMT and CTT is the necessity of variation in the distribution of items, i.e. the bandwidth of the scale. In the classical paradigm, typically no attention is paid to the issue of bandwidth, whereas the Rasch model asks for markedly different item locations. For many constructs it may be challenging to generate items that potentially do provide such variation.

In our comparative study of a scale measuring technophobia was investigated in three countries (the UK, Austria, and Mexico). The results of the MG-CFA and the Rasch approach were very similar to a large extent. Contrary to what we expected, the Rasch model retained more items than the classical approach. Still, we claim that, in general, the CFA would keep more items – although wrongly from a Rasch point of view. The reason is the higher flexibility of CFA in terms of item discrimination. In our case, the data meet stringent levels of equivalence. Metric invariance holds for all items in the final scale and half of the loadings can even be constrained across items. Thus, it is not surprising that the data fit the Rasch model reasonably well. On the other hand, the Rasch model eliminated one item that seems to be doubtful but nevertheless reaches even scalar invariance in the MG-CFA. The ease with which the Rasch model handles missing data allows us to split items affected by misfit in one of the groups and discard the responses in that group only. Consequently, more items, specific to only one or two groups can be retained enhancing the precision of the person measures. More pronounced correlations with the external construct of innovativeness support this conclusion.

The use of real data may be seen as an inherent limitation of the present study but also as an advantage because only real data allow for investigating the behaviour of the models in a real world situation of true responses rather than generated responses following a particular model. We claim that a conclusive evaluation of the methods discussed requires reference to theory. If one still wants to draw conclusions purely from empirical applications, one study is certainly insufficient. A meta-study would help identify persistent patterns. Unfortunately, up to now, applications of the Rasch model in single culture studies are scarce, let alone in multi-cultural settings with a full comparison of the MG-CFA and the Rasch approach.

In terms of validating scales based on the Rasch approach, only the adoption of the Rasch philosophy by the researcher, in particular during item generation, may help fully exploit the potential of the Rasch model. Significant variation of the item location would, in all likelihood, cause more prominent differences in scale validation between CFA and the Rasch model. The justification of linear, interval-scaled measures of latent constructs in marketing would certainly benefit from a more widespread application of Rasch analyses, in particular in cross-cultural studies but also whenever the measures of individuals are of interest.

References

Andrich, D. (1978), "A rating formulation for ordered response categories", *Psychometrika*, Vol. 43 No. 4, pp. 561-73.

Andrich, D. (1988), "A general form of Rasch's extended logistic model for partial credit scoring", *Applied Measurement in Education*, Vol. 1 No. 4, pp. 363-78.

Andrich, D., Sheridan, B.S. and Luo, G. (2003a), "Displaying the Rumm2020 analysis", working paper, RUMM Laboratory, Perth.

Andrich, D., Sheridan, B.S. and Luo, G. (2003b), *Rumm2020: Rasch unidimensional Measurement Models*, RUMM Laboratory, Perth.

Baker, F.B. (1994), *Genir V: Computer Program for Generating Item Response Theory Data*, University of Wisconsin, Laboratory of Experimental Design, Wisconsin, MA.

Baumgartner, H. and Steenkamp, J-B.E.M. (1998), "Multi-group latent variable models for varying numbers of items and factors with cross-national and longitudinal applications", *Marketing Letters*, Vol. 9 No. 1, pp. 21-35.

Boomsma, A. and Hoogland, J.J. (2001), "The robustness of Lisrel modeling revisited", in Cudeck, R., Dau Toit, S. and Sörbom, D. (Eds), *Structural Equation Modeling: Present and Future, a Festschrift in Honor of Karl Jöreskog*, SSI Scientific Software International, Lincolnwood, IL, pp. 139-68.

Brod, C. (1984), *Technostress: The Human Cost of the Computer Revolution*, Addison-Wesley, Reading, MA.

Cheung, G.W. and Rensvold, R.B. (1998), "Cross-cultural comparisons using non-invariant measurement items", *Applied Behavioral Science Review*, Vol. 6 No. 1, pp. 93-110.

Churchill, G.A. (1979), "A paradigm for developing better measures of marketing constructs", *Journal of Marketing Research*, Vol. 16 No. 1, pp. 64-73.

Craig, C.S. and Douglas, S.P. (2005), *International Marketing Research*, 3rd ed., Wiley, Chichester.

Davis, H.L., Douglas, S.P. and Silk, A.J. (1981), "Measure unreliability: a hidden threat to cross-national marketing research?", *Journal of Marketing*, Vol. 45 No. 1, pp. 98-109.

Diamantopoulos, A. (2005), "The C-OAR-SE procedure for scale development in marketing: a comment", *International Journal of Research in Marketing*, Vol. 22 No. 1, pp. 1-9.

Embretson, S.E. and Reise, S.P. (2000), *Item Response Theory for Psychologists*, Lawrence Erlbaum Associates, Mahwah, NJ.

Ewing, M.T., Thomas, S. and Sinkovics, R.R. (2005), "An alternate approach to assessing cross-cultural measurement equivalence in advertising research", *Journal of Advertising*, Vol. 34 No. 1, pp. 17-36.

Fan, X. (1998), "Item response theory and classical test theory: an empirical comparison of their item/person", *Educational and Psychological Measurement*, Vol. 58 No. 3, p. 357.

Finn, A. and Ujwal, K. (2005), "How fine is C-OAR-SE? A generalizability theory perspective on Rossiter's procedure", *International Journal of Research in Marketing*, Vol. 22 No. 1, pp. 11-21.

Hirschman, E.C. (1980), "Innovativeness novelty seeking, and consumer creativity", *Journal of Consumer Research*, Vol. 7 No. 3, pp. 283-95.

Jaufmann, D. (1991), *Einstellungen Zum Technischen Fortschritt: Technikakzeptanz Im Nationalen und Internationalen Vergleich*, Campus Verlag, Frankfurt/Main.

Jöreskog, K.G. and Dag, S. (2003), *Lisrel 8.54*, Scientific Software International Inc., Chicago, IL.

Lord, F.M. and Novick, M.R. (1968), *Statistical Theories of Mental Test Scores*, The Addison-Wesley Series in Behavioral Science: Quantitative Methods, Addison-Wesley, Reading, MA.

Masters, G.N. (1982), "A Rasch model for partial credit scoring", *Psychometrika*, Vol. 47 No. 2, pp. 149-74.

Meade, A.W. and Lautenschlager, G.J. (2004), "A comparison of item response theory and confirmatory factor analytic methodologies for establishing measurement equivalence/invariance", *Organizational Research Methods*, Vol. 7 No. 4, pp. 361-88.

Meier, S.T. (1985), "Computer aversion", *Computers in Human Behavior*, Vol. 1 No. 2, pp. 171-9.

Mullen, M.R. (1995), "Diagnosing measurement equivalence in cross-national research", *Journal of International Business Studies*, Vol. 26 No. 3, pp. 573-96.

Price, L.L. and Ridgeway, N.M. (1983), "Development of a scale to measure use innovativeness", in Richard, P.B. and Tybout, A.M. (Eds), *Advances in Consumer Research*,Vol. 10, Association for Consumer Research, Ann Arbor, MI, pp. 679-84.

Rasch, G. (1961), "On general laws and the meaning of measurement in psychology", paper presented at Berkeley Symposium on Mathematical Statistics and Theory of Probability, University of California Press, Berkeley, CA,Vol. IV, pp. 321-33.

Rasch, G. (1977), "On specific objectivity. An attempt at formalizing the request for generality and validity of scientific statements", *Danish Yearbook of Philosophy*, Vol. 14, pp. 58-94.

Rasch, G. (1980), *Probabilistic Models for Some Intelligence and Attainment Tests*, Mesa Press, Chicago, IL, reprint of 1960, Danish Institute of Educational Research.

Raub, A.C. (1982), "Correlates of computer anxiety in college students", dissertation, University of Pennsylvania, Philadelphia, PA.

Reise, S.P., Widaman, K.F. and Pugh, R.H (1993), "Confirmatory factor analysis and item response theory: two approaches for exploring measurement invariance", *Psychological Bulletin*, Vol. 114 No. 3, pp. 552-66.

Rentz, J.O. (1987), "Generalizability theory: a comprehensive method for assessing and improving the dependability of marketing measures", *Journal of Marketing Research*, Vol. 24 No. 1, pp. 19-28.

Röglin, H.C. (1994), *Technikängste und Wie Man Damit Umgeht*, VDI-Verlag, Düsseldorf.

Rosen, L.D. and Weil, M.M. (1990a), "Computers classroom instruction, and the computerphobic university student", *Collegiate Microcomputer*, Vol. 8 No. 4, pp. 275-83.

Rosen, L.D. and Weil, M.M. (1990b), "Myths and realities of computerphobia", *Anxiety Research*, No. 3, pp. 175-91.

Rossiter, J.R. (2002), "The C-OAR-SE procedure for scale development in marketing", *International Journal of Research in Marketing*, Vol. 19 No. 4, pp. 305-35.

Rossiter, J.R. (2005), "Reminder: a horse is a horse", *International Journal of Research in Marketing*, Vol. 22 No. 1, pp. 23-5.

Salzberger, T., Sinkovics, R.R. and Schlegelmilch, B.B. (1999), "Data equivalence in cross-cultural research: a comparison of classical test theory and latent trait theory based approaches", *Australasian Marketing Journal*, Vol. 7 No. 2, pp. 23-38.

Schaffer, B.S. and Riordan, C.M. (2003), "A review of cross-cultural methodologies for organizational research: a best-practices approach", *Organizational Research Methods*, Vol. 6 No. 2, pp. 169-215.

Singh, J. (1995), "Measurement issues in cross-national research", *Journal of International Business Studies*, Vol. 26 No. 3, pp. 597-619.

Singh, J. (2004), "Tackling measurement problems with item response theory: principles, characteristics, and assessment, with an illustrative example", *Journal of Business Research*, Vol. 57 No. 2, pp. 184-208.

Sinkovics, R.R., Stöttinger, B., Schlegelmilch, B.B. and Ram, S. (2002), "Reluctance to use technology-related products: development of a technophobia scale", *Thunderbird International Business Review*, Vol. 44 No. 4, pp. 477-94.

Steenkamp, J-B. and Baumgartner, H. (1998), "Assessing measurement invariance in cross-national consumer research", *Journal of Consumer Research*, Vol. 25 No. 1, pp. 78-90.

Tabachnick, B.G. and Fidell, L.S. (2001), *Using Multivariate Statistics*, 4th ed., Harper Collins College Publishers, New York, NY.

Vandenberg, R.J. (2002), "Toward a further understanding of and improvement in measurement invariance methods and procedures", *Organizational Research Methods*, Vol. 5 No. 2, pp. 139-58.

van de Vijer, F. and Leung, K. (1997), *Methods and Data Analysis for Cross-Cultural Research*, 1st ed., Sage, Thousand Oaks, CA.

Appendix

Code	Item retained in the final MG-CFA model	Item retained in at least one group in the final Rasch model	Item wording
v01		X	I feel some anxiety when I approach an ATM
v02			I prefer to have people handle my bank activities than to use an ATM
v03			ATMs are fun to use
v04			I feel comfortable when using ATMs
v05			I want to learn more about using ATMs
v06	X	X	I worry about making mistakes when using ATMs
v07	X	X	Using ATMs is time-consuming
v08		X	ATMs agitate me
v09		X	It takes me a long time to complete bank transactions when using an ATM
v10			I think most people know how to use ATMs better than I
v11			I resent that ATMs are becoming so prevalent in our daily lives
v12			I can conduct my bank transactions without using an ATM
v13		X	I refuse to use ATMs
v14	X		It is easy to learn how to use ATMs
v15			I feel frustrated when I use an ATM
v16			I feel inadequate about my ability to use ATMs
v17		X	Thinking about ATMs makes me nervous
v18	X	X	I do not trust ATMs with my money
v19			ATMs make things too complicated
v20		X	ATMs are intimidated
v21			I wish I were more adept at using ATMs
v22			ATMs make bank transactions easier
v23		X	ATMs seem very complicated

(*continued*)

Table AI.
Item pool of the technophobia scale

Code	Item retained in the final MG-CFA model	Item retained in at least one group in the final Rasch model	Item wording
v24			I like that ATMs are so convenient
v25			I feel more confident dealing with a human teller than an ATM
v26		X	Machines should not handle people's money transactions
v27	X	X	I find ATMs instructions confusing
v28			I have no fear of ATMs
v29	X	X	I feel confident that I could teach someone how to use an ATM
v30			I do not go to the bank after lobby and drive-thru teller hours

Source: Sinkovics *et al.* (2002)

Table AI.

Corresponding author
Thomas Salzberger can be contacted at: thomas.salzberger@wu-wien.ac.at