Partial Credit Models for Scale Construction in Hedonic Information Systems

Mair, Patrick; Treiblmaier, Horst

Published: 01/01/2008

Document Version
Publisher's PDF, also known as Version of record

Partial Credit Models for Scale Construction in Hedonic Information Systems

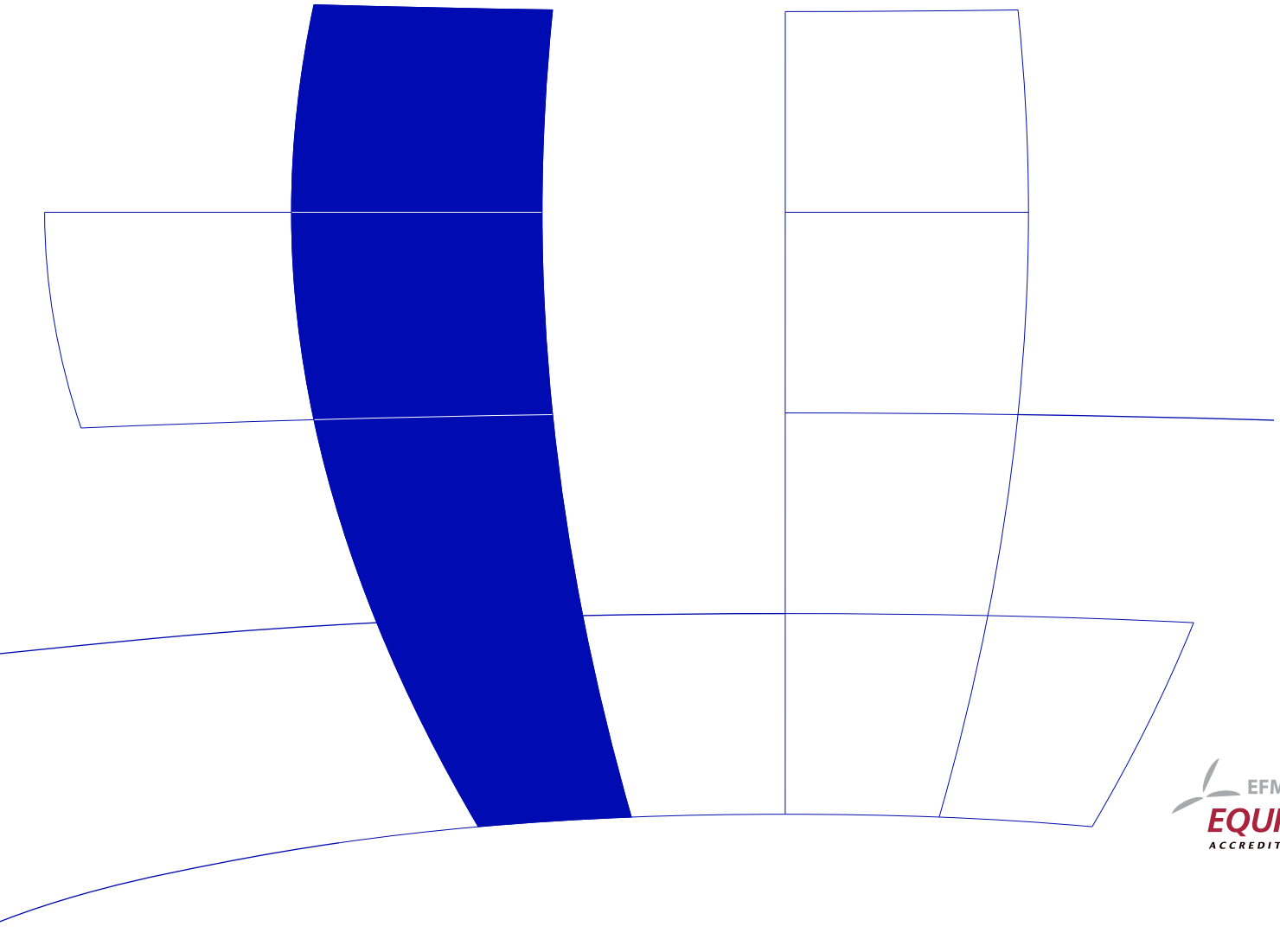
Patrick Mair, Horst Treiblmaier

Department of Statistics and Mathematics
Wirtschaftsuniversität Wien

Research Report Series

Report 62
March 2008

http://statmath.wu-wien.ac.at/



Partial Credit Models for Scale Construction in Hedonic Information Systems

Patrick Mair
Department of Statistics and Mathematics
Wirtschaftsuniversität Wien

Horst Treiblmaier
Department of Information Systems and Operations
Wirtschaftsuniversität Wien

Abstract

Information Systems (IS) research frequently uses survey data to measure the interplay between technological systems and human beings. Researchers have developed sophisticated procedures to build and validate multi-item scales that measure real world phenomena (latent constructs). Most studies use the so-called classical test theory (CTT), which suffers from several shortcomings. We first compare CTT to Item Response Theory (IRT) and subsequently apply a Rasch model approach to measure hedonic aspects of websites. The results not only show which attributes are best suited for scaling hedonic information systems, but also introduce IRT as a viable substitute that overcomes severall shortcomings of CTT.

1 Introduction

Over the last couple of years, social science research in general and Information Systems (IS) research in particular has been dominated by empirical papers that use survey data either to create new measurement scales (Webster and Martocchio, 1992; Salisbury et al., 2002), or to apply previously validated scales to measure constructs that can subsequently be used to test hypotheses and theories (Weber, 2003). Additionally, new research is being published in IS journals on how to improve current methods. In most cases, the authors rely on fundamental principles that have been developed and refined in classical test theory (CTT) over the last decades. They discuss issues such as the relation between a construct and its items (Petter et al., 2007), moderation errors (Carte and Russell, 2003), generalizability (Lee and Baskerville, 2003), and how to find new approaches to model latent variables (Chin and Marcolin, 2003). However, the underlying measurement theory has not been questioned.

In a recent research note, Allport and Kerler (2003, p. 356) acknowledge that "measurement is perhaps the most difficult aspect of behavioral research". Accordingly, researchers have developed alternative approaches on how to measure latent constructs in social science research. (Stevens, 1946) gave the classical definition of measurement. He points out that measurement is the assignment of numerals to events or objects according to rules. This notion implies a very wide interpretation of this term, and consequently has been criticized over the last decades, as for instance by Michell (1999). Mitchell defines "measurement as the discovery or estimation of the ratio of a magnitude of a quantity to a unit of the same quantity" (p. 222, see also Salzberger, 2007). *Quantity* is an attribute possessing ordinal and additive structure, whereas *quantification* is the corresponding process of showing that an attribute is quantitative, and Mitchell devises procedures to measure it. Finally, he defines a *unit* as a specific magnitude of a quantity relative to which measurements are made.

Psychometricians such as Thurstone (1925) and Rasch (1960) have formulated statistical models to achieve the *objective measurement* of latent traits. Rasch proposed a probabilistic model known as *Rasch model* (Rasch, 1960). His model allows researchers to link items to the

trait they are supposed to measure. Even though the first IRT models were introduced decades ago, their application in scholarly research is still limited (Borsboom, 2006). Today, most research papers utilizing IRT can be found in psychology, and at the same time these papers are slowly gaining popularity in marketing research. Comparatively few research papers that use the Rasch model have been published in leading IS journals (notable exemptions include e.g. Dekleva and Drehmer (1997) and Alvarez et al. (2007)). However, in recent years several publications have clearly shown the advantages of this measurement approach and thus have sparked new interest (Salzberger and Sinkovics, 2004). Additionally, the current PISA study (Programme for International Student Assessment), which is conducted in more than 60 OECD member countries so far (OECD, 2007), has successfully applied an extended version of the Rasch model (Adams et al., 2007).

Typically, when researchers measure latent variable(s), they strive to find a "good" subset of items which allows for a reliable measurement of the underlying construct. However, objective *measurement* is based on fundamental requirements which, as we show in the next sections, cannot be accomplished by the approaches that are commonly used. In contrast, we illustrate how IRT models can be applied to achieve high-quality measurement by means of an objective measurement of a hedonic IS system.

2 Objective Measurement Using IRT models

2.1 Classical Test Theory

The common approach in scale construction is known as *classical test theory* (CTT, Lord and Novick, 1968). Its basic equation is $X = T + E$. In this basic equation, with X as the observed score, T as the true score, and E as the measurement error, the right-hand side is completely unknown, such that to meet the equation T and E can be chosen arbitrarily. Thus, this equation a tautology rather than a statistical model (Fischer, 1974). In spite of this fact, researchers usually compute reliability coefficients based on this basic expression. Reliability is defined as $\rho^2(X,T) = \sigma^2(T)/\sigma^2(X)$. Since T is unknown, we cannot compute its variance $\sigma^2(T)$. As a consequence, we need additional assumptions in terms of measurement equivalence of test splitting, so reliability is commonly estimated as internal consistency by means of Cronbach's α (Cronbach, 1951). Researchers make their item selection by some rules of thumb (e.g. $\alpha > .70$) without the possibility of testing the results in a statistical manner.

The whole process of item selection is based on correlation coefficients. The square root of the reliability is expressed as $r(X,T)$ and the discriminatory power of item i (i.e. whether item i measures "something near identical" than the test composite score) as $r(X_i, X)$. Items that are highly correlated are retained and items that are weakly correlated with other items are eliminated. Generally, correlations are sample dependent, which implies different test reliabilities for homogeneous and heterogeneous samples, respectively (Fischer, 1974).

CTT does not provide the terminology of a *latent construct* or *latent trait* Θ, since there is no underlying theory implied on how "measurement" is achieved. Basically, researchers treat the person/item sum scores, which are based on ordinal indicators and therefore still on an ordinal scaling level, as if they were interval scaled and consider the sum scores as "estimates" of person ability and item difficulty. Obviously, there is a fundamental measurement problem. Moreover, to be allowed to sum up the item/person scores, the construct under investigation has to be unidimensional, and such constructs cannot be tested within a CTT framework. In general, CTT does not allow for statistical model testing, and hence, excluding items follows rules of thumb.

CTT has been intensely criticized over the last decades, mainly in the field of psychology (e.g. Fischer, 1974; Weiss and Davison, 1981; Hambleton and Jones, 1993; Borsboom, 2006) but also within a marketing context (Salzberger, 2007). In marketing, researchers refer mainly to Churchill's measurement paradigm (Churchill, 1979), which is founded on CTT. However, in spite of these obvious shortcomings, CTT prevails when researchers think of scale construction.

2.2 Reflective and Formative Scale Construction

The base model of classical test theory translates straightforwardly to *confirmatory factor analysis* (CFA):

$$\mathbf{X} = \mathbf{\Lambda}\boldsymbol{\xi} + \boldsymbol{\delta}. \tag{1}$$

The data we observe in the matrix $\mathbf{X}$ are explained by a latent variable $\boldsymbol{\xi}$ with the corresponding regression coefficients $\mathbf{\Lambda}$ (called *loadings*) and the error term $\boldsymbol{\delta}$. Obviously, CFA poses a linear relation between $\boldsymbol{\xi}$ and $\mathbf{X}$. We refer to this approach as a *reflective* method for scale construction in which $\boldsymbol{\xi}$ causes $\mathbf{X}$. CFA assumes a metric scale level of the indicators. However, for test items, IS researchers do not use metric scales, but frequently use Likert scales. Furthermore, normality assumptions must be imposed that allow the application of tetrachoric or polychoric correlation coefficients. The computations in CFA are performed at an (aggregated) correlation level, which implies loss of information as compared to analyzes based on the observed response patterns (Salzberger and Sinkovics, 2004).

Recently, *formative* scale construction has received growing interest in social science literature (e.g. Diamantopoulos and Winklhofer, 2001; Petter et al., 2007). Compared to CFA, formative scales reverse the causal relation such that $\mathbf{X}$ causes $\boldsymbol{\xi}$. Hence, equation 1 changes to

$$\boldsymbol{\xi} = \mathbf{X}\boldsymbol{\beta}, \tag{2}$$

where $\boldsymbol{\beta}$ are weights. In formative scaling we calculate a weighted sum of indicators,i.e., $\boldsymbol{\xi}$ is not a latent variable that is measured, but rather an index. This approach corresponds to a *principal component analysis* (PCA). If K is the number of variables, then PCA reduces the K-dimensional space $\mathbb{R}^K$ to a lower-dimensional space $\mathbb{R}^m$ where $m \ll K$. Since there are no underlying latent traits, we do not consider it as measurement model in a strict sense (Salzberger, 2007).

2.3 IRT Measurement

As was noted earlier, when the goal is to *measure* a latent construct, CTT and related methods lead to severe problems. Borsboom (2006, p.429) quotes that "in an alternative world, where CTT was never invented, the first thing a researcher, who has proposed a measure for a theoretical attribute, would do is to spell out the nature and the form of the relationship between the attribute and its putative measures". That is exactly what IRT does by overcoming several limitations of CTT.

First, the linear relation between indicators and a categorical response, which is assumed in CTT, is usually not feasible. Instead, S-shaped relations seem more realistic, i.e., logit or probit functions or complete nonparametric step functions. Second, the analysis should be carried out on a response pattern level where the researcher can use the full amount of available information, rather than on an aggregated correlation level. In particular cases, it makes sense to use the sum scores as a sufficient statistic (i.e., in Rasch measurement). Third, items should not be selected following some rules of thumb on underlying approximative and sample dependent measures. Instead, the researcher should choose the items based on statistical tests, and do so without worrying about artificial concepts such as reliability, internal consistency and construct validity. Fourth, it should be possible, as is the common procedure in statistics, to carry out goodness-of-fit tests for the whole model.

It would also be useful to obtain detailed information at an item and person level simultaneously. Each item i and each person v should be assigned a parameter (i.e., difficulty β_i and ability θ_v) that would allow for a probabilistic analysis of the response behavior. Item and person parameters should be on an interval scale, which would make possible the interpretation of distances between items and persons on Θ. If the items and persons are on the same scale, then statements about the response probability of person v on item i can be achieved. The final item subset should be homogeneous in terms of the trait that the items measure, and heterogeneous in terms of their difficulty, i.e., they should allow to map persons for a wide range of abilities.

IRT offer all of these options, which, by means of β_i and θ_v, can be described as follows. The base of analysis is a $(0, 1)$ persons $\times$ items data matrix $\mathbf{X}$ of dimension $N \times K$. Item response patters $\mathbf{x}_i$ and person response patterns $\mathbf{x}_\mathbf{v}$ are indicators for β_i and θ_v. Other than in CFA, neither causal nor distributional assumptions need to be imposed. The patterns

$\mathbf{x}_i$ and $\mathbf{x}_v$ are still on an ordinal level, but β_i and θ_v are on an interval scale. The basic functional relation is $P(\mathbf{X}_{vi} = x_{vi}) = f(\beta_i, \theta_v)$. Different IRT approaches exist in terms of choosing the function f (e.g., logistic) and in terms of the number of item-related parameters. For instance, in addition to the difficulties β_i, the researcher might wish to allow for item-discrimination parameters α_i or guessing parameters γ_i, which in some situations might be more realistic. Depending on the degree of parameterization and the overall goal of the analysis, two conceptual approaches in IRT exist:

- *Item selection approach*: The aim is to find "high-quality" items in terms of fairness, sample independence, discrimination, and heterogeneous difficulties. The corresponding models, called Rasch (or Rasch-type) models, are parsimonious in terms of the item-related parameters. These models allow for *objective measurement*.
- *Modeling approach*: If the researcher is not primarily interested in selecting items in a very restrictive manner, but instead wishes to analyze a person's response behavior, then he or she should take into account higher parameterized models or models with covariates.

IRT models that follow the second approach overcome the shortcomings of CTT and related methods, but the measurement (items, persons) is not *objective*. Rasch (1960) reasoned on requirements to be fulfilled such that a specific proposition can be regarded as "scientific". His conclusion was that a basic requirement is the *objectivity of comparisons* (Rasch, 1961) and he formulated the epistemological theory of *specific objectivity* (SO): *objective* because any comparison of a pair of parameters (items/persons) should be independent of any other parameters or comparisons; *specifically objective* because the comparison made was relative to some specified frame of reference (Andrich, 1988). In other words, under SO, two persons v and w with abilities θ_v and θ_w are comparable independently from the remaining persons in the sample and independently from the item subset with which they are presented. In turn, two items i and j with β_i and β_j are comparable independently from the remaining items in the subset and independently from the persons in the sample (Mair and Hatzinger, 2007c).

The strict requirements of Rasch (or Rasch-type) models, demand that these models be considered as a general "seal of approval" for tests and scales, respectively.

3 Rasch Measurement in Social Science Research

3.1 Properties of Rasch Models

In his groundbreaking work, Rasch (1960) presented a probabilistic model that could be used to study the response behavior of individuals on dichotomous items. It poses a logistic relation between the ability θ_v of a person v and the probability for a correct response on item i. Each item gets a difficulty parameter β_i. The formal representation, which is known as *Rasch model*, is

$$P(\mathbf{X}_{vi} = 1) = \frac{exp(\theta_v - \beta_i)}{1 + exp(\theta_v - \beta_i)} \tag{3}$$

This model has some remarkable features. We describe them briefly, since they also apply for polytomous Rasch models, that we use for the subsequent analysis. As we mentioned in the last section, SO is the necessary condition for objective measurement. The "mathematical translation" of SO in terms of parameter estimation is the *conditional maximum likelihood* (CML) approach. The Rasch model assumes that the person raw score r_v is a sufficient statistic. That implies that we do not need to know the exact $(0, 1)$-pattern of a person, but that all the information needed for parameter estimation is contained in r_v. Thus, the Rasch model forms an *exponential family* and the log-likelihood is established by conditioning on the raw score, i.e. $\log L_C(\hat{\boldsymbol{\beta}}; \mathbf{X})$.

We can show that $\boldsymbol{\theta}$ disappears from the likelihood equation, which implies that the item parameters can be estimated independently from the person parameters. This property is called *parameter separability*. Besides having the mathematical formulation of SO, the parameter separation circumvents another technical problem. Person parameters $\boldsymbol{\theta}$ are *nuisance parameters*, i.e., the larger the sample size, the more parameters we have to estimate. But,

if we calculate the parameters simultaneously, as is done in *joint maximum likelihood* (JML), leads to inconsistent estimates (see, e.g., Haberman, 1977).

We note that a further common and very flexible approach to estimating IRT models is *marginal maximum likelihood* (MML; Bock and Aitkin, 1981) where the θ's are assumed to follow a certain distribution (e.g. a standard normal). If this assumption is violated, then the parameters can be biased. In addition, the concept of *person-free item assessment* (which is based on SO) no longer holds. However, there are nonparametric approaches that turn out to be equivalent to CML (de Leeuw and Verhelst, 1986). Detailed explanations on parameter estimation can be found in Baker and Kim (2004).

Another implication of SO and CML estimation, respectively, is the *sample invariance*. If the Rasch model holds, then the item parameters are invariant over arbitrary person subgroups. For instance, we can split the sample by internal factors (e.g., raw score mean, median, random) or external factors (e.g., grouping variables such as gender and culture group), and estimate the parameters for each subgroup. Due to the above-mentioned sample invariance, the item parameters must be the same across subgroups. We can perform statistical tests based on this property, both item-wise (Wald-test) and as goodness-of-fit tests. As model test we use the likelihood-ratio (LR) test proposed by Andersen (1973), which is based on (person) sample splits. We perform item selection by means of residual-based itemfit statistics (Smith, 2004).

In addition to the properties we mention above, Rasch models have further assumptions. First, they are models of *unidimensional* scaling. Thus, they allow for only one underlying latent construct. Second, *local independence* is necessary which assumes conditional independence of the item responses, i.e. $p(\mathbf{x}_v|\theta_v) = \prod_{i+1}^{K} p(x_{vi}|\theta_v)$. Third, Rasch models do not allow the logistic curves determined by Equation 3 (*item characteristic curves*) to cross. Due to the item selection process and to achieve model fit, we eliminate items that contradict at least one of these assumptions. Those items that remain in the final homogeneous item subset measure the latent construct in an objective manner.

3.2 Polytomous Rasch Models

In many practical situations dichotomous item responses are too restrictive. That is especially true in social science research, where Likert-scales are commonly used for assessing individuals' attributes. For corresponding polytomous items Rasch (1961) proposed an extension of his classical model. We note that all the explanations for the simple Rasch model in terms of SO, CML estimation, and model testing in the former section apply equally well for polytomous Rasch models. Based on Rasch's polytomous expression, Andersen (1995) gives the following representation. The probability of a response h ($h = 0, \ldots, m_i$ where m_i is the number of categories -1 on item i) is given by

$$P(X_{vi} = h) = \frac{\exp(h\theta_h + \beta_{ih})}{\sum_{l=0}^{m_i} \exp(l\theta_v + \beta_{il})}. \tag{4}$$

This representation is known as the *partial credit model* (PCM; Masters, 1982) with β_{ih} as item-category parameters. A more restrictive model is the *rating scale model* (RSM; Andrich, 1978), which decomposes β_{ih} into an item parameter and a category parameter. We note that for the RSM, the number of response categories must be the same for each item.

Among other models (see Fischer and Molenaar, 1995) researchers consider these approaches as Rasch models. Since the raw score is sufficient for analysis, we can apply CML estimation and these models conform to the theory of SO and objective measurement. Various other polytomous IRT extensions have been proposed (for an overview see van der Linden and Hambleton, 1997). However, most of them can not be considered as Rasch-type models, so they are better suited for *modeling* items responses rather than for the *selection* of items to achieve objective measurement.

Since we focus on the item selection approach, we limit further explanations to the PCM. The RSM is too restrictive, since it requires equal distances between adjacent response categories of the measurement scale. Both models are commonly referred to as *adjacent-categories logits model* (Tuerlinckx and Wang, 2004): Basically, it estimates log-odds for a certain category h with respect to category $h - 1$.

An important issue is the interpretation of the item-category parameters β_{ih}. These parameters are often transformed into *category intersection parameters* δ_{ij} with $j = 0 \ldots m_i$

as follows: Originally, Masters (1982) formulated Equation 4 in terms of $\beta_{ih} = -\sum_{j=0}^{h} \delta_{ij}$, i.e.,

$$P(X_{vi} = h) = \frac{\exp \sum_{j=0}^{h} (\theta_v - \delta_{ij})}{\sum_{l=0}^{m_i} \exp \sum_{j=0}^{l} (\theta_v - \delta_{ij})}. \quad (5)$$

If we estimate the PCM by Equation 4, then the item-categories are converted into intersection parameters as follows: $\delta_{i0} = -\beta_{i0}$; $\delta_{i1} = \beta_{i0} - \beta_{i1}$; $\delta_{i2} = \beta_{i1} - \beta_{i2}$ etc. In general, polytomous IRT models allow for different parameterizations, but eventually, the intersection parameters are straightforward to interpret. The parameters δ_{ij} refer to the points on the latent trait where the *item category curves* (ICC) intersect. Based on these intersection parameters, we can compute item location parameters ν_i in terms of $\nu_i = m_i^{-1} \sum_{j=0}^{m_i} \delta_{ij}$. These parameters can be interpreted as general difficulty parameters. (Embretson and Reise, 2008) give an extensive discussion of parameter interpretations and relations between polytomous IRT models.

Within the context of item selection, the main focus is on the item(-category) parameters that we can estimate independently from person parameters. In the current example we are not interested in θ-estimation, since the websites are rated by individuals. Our aim is to establish a homogeneous subset of items that allows for an objective measurement of hedonism. We estimate the $\hat{\boldsymbol{\theta}}$ for item selection, since we use item-fit statistics based on Pearson residuals that are approximately standard normal distributed. To compute the model probabilities $P(X_{vi})$, we require the θ's, which we estimate by ordinary ML in which we include the CML-based item parameters into the likelihood equation (see e.g. Hoijtink and Boomsma, 1995).

3.3 Item Interpretation

So far, we have made all elaborations by using the classical psychometric terms: "item difficulties" and "person abilities". The more to the left on the latent continuum, the less difficult/able is an item/person, the more to the right, the more difficult/able is a item/person. However, IRT models are not limited to cognitive ability testing. Whenever researchers have a set of items that are supposed to be measures of a single (interval scaled) latent continuum Θ, they can apply IRT models. We note that the interval scale implies that Θ must have two directions; roughly speaking "more" and "less".

Next, we illustrate how to measure hedonic information systems exemplified by attributes of websites. Of crucial importance is the interpretation of the scale direction, which differs from the interpretation of items that measure ability. It is straightforward to see that moving to the right on Θ means "more hedonism" and moving to the left stands for "less hedonism". As in many other non-ability constructs (see e.g. Salzberger (2007) for various marketing scales) the $-\infty$-direction does not refer to the opposite of the $+\infty$-direction, i.e., the opposite of hedonism. Instead, the extreme point at $-\infty$ refers to "null or no hedonism". We note that items with higher item parameters are not better in some sense than items with lower parameters. If items are Rasch-homogeneous, all of them share high-quality properties, regardless of whether they lie closer to $+\infty$ or $-\infty$. To score persons, a reliable ability test should have a wide range of items in terms of their difficulty. The same assumption applies to our website hedonism scale.

To introduce the readers to the basic concepts of IRT, we have used the classical psychometric terminology (i.e., item difficulties and person abilities) in sections above, but in the remainder we use *subject location* and *item location* instead, which better reflect the focus of our research (Salzberger (1999) proposes the term *item affectivity*).

4 Measuring Hedonic Information Systems

4.1 The concept of Hedonism in IS Research

Previous research shows the importance of hedonism for the usage of information systems (van der Heijden, 2004). Instead of seeing individuals as rational beings who actively process huge amounts of information before making shopping decisions (cf. Venkatraman and MacInnis, 1985), researchers such as Hirschman and Holbrook (1982) and Holbrook and Hirschman (1982) highlight the hedonic, esthetic and symbolic nature of the consumption

process. Generally speaking, the difference between utilitarian and hedonic behavior can be seen as performing an act "because you love it" as opposed to "getting something" (Triandis, 1977).

With the advent of technical systems that offer advanced multimedia capabilities, and particularly the World Wide Web, utilitarian and hedonic concepts have begun to intermingle. Besides offering interesting and informative content, websites must appeal to users' hedonic predispositions in order to ensure an entertaining online shopping experience (Huang, 2005). Additionally, websites use features such as online games, e-cards, wallpapers, sweepstakes, or, even more sophisticated, virtual communities to make users linger at their site, revisit it (Cotte et al., 2006), or to increase involvement with a certain product or brand (Füller et al., 2006). In the online world, the traditional boundaries between utilitarism and hedonism become blurred.

Sometimes the constructs used in IS literature to describe non-rational behavior are used interchangeably. For example, (Cheung et al., 2000, p. 3), use the notion of affect. They describe affect as the "emotional response to the thought of a behavior", which are "feelings of joy, elation, or pleasure, or depression, disgust, displeasure, or hate associated by an individual with a particular act" (Triandis, 1980). Consequently, van der Heijden (2004) adopts these four items for the measurement of "perceived enjoyment". Lin et al. (2005) use parts of a scale from Moon and Kim (2001) to measure "playfulness". Their items (e.g., "When interacting with the web portal, I am not aware of the time as it elapses") seemingly overlap with the concept of flow, introduced by Csikszentmihalyi (1990). Other constructs that are associated with hedonism and are frequently used in IS studies include cognitive absorption (Agarwal and Karahanna, 2000), playfulness (Webster and Martocchio, 1992) and enjoyment (Davis et al., 1992). For all of these constructs, activities are either performed "per se" or the individual gets completely involved in the activity. For the purpose of this research, we initially intend to create a item base (i.e., the attributes of website) that is as broad as possible, which we can later refine.

4.2 Data Description

We used several steps to collect and clean the data. To ensure that the attributes represent all facets of the concept under investigation (i.e., content validity, Straub (1989)), we used a panel of seven experts to generate a list of properties that are important for customer portal websites. We designed this phase as a brainstorming session, with the major objective of coming up with as many attributes as possible without any evaluation or rating. The experts produced a total of 79 items. Subsequently, we used the same panel of experts to group the items they came up with and to filter out synonyms. They created a total of five groups, three of which describe hedonic aspects ("Games, Fun and Dynamics", "Emotion", and "Static Design Aspects"). After performing ten preliminary tests to ensure the understandability of the items, we conducted an online survey in which a convenience sample of 291 Internet users rated the importance of those 25 attributes for measuring hedonic concepts. We used a scale with a range from zero ("not important") to four ("very important") to assess the significance of the single attributes. Therefore, the data matrix $\mathbf{X}$, which we use for all subsequent analyses, consists of 291 subjects and 25 items. Since we conducted all surveys in German, we used a translation and back-translation approach to ensure semantic consistency. The original items were translated by the authors and then back-translated by independent translators. When no agreement could be reached, we consulted another independent translator.

4.3 Stepwise Item Elimination and Final Item Subset

We perform all computations with the `eRm` package (Mair and Hatzinger, 2007a,c) in `R` (R Development Core Team, 2007), which uses CML estimation and allows for computation of the test statistics described in Section 3.1. To achieve a final set of items, we used the following steps:

1. Estimate item and subject parameters of the PCM.
2. Compute itemfit statistics based on residuals.
3. Eliminate the item with the smallest p-value.

4. Compute LR-test for different person subsplits: If LR is significant, go back to step (1) and proceed with item elimination. Otherwise, the procedure stops and we obtain the final model.

Unless the data do not fit the PCM, we eliminate items successively and re-fit the model. At the end, we get a set of homogeneous items that comply with the restrictive Rasch criteria (i.e., they are Rasch-homogeneous). The reason for fitting the LR-test after each step is that this statistic, which is a global model test, proves the model fit as a whole. Item-fit statistics are residual based and compare a theoretical probability with an observed integer value. Thus, this criterion is only suitable for indicating which items should be eliminated. It is not suited to testing for model fit.

We start our analysis with a total set of 25 items which might be suitable for measuring the hedonic aspects of a website. We eliminate the following items in this exact order: plain, suitable for children, interactive, customized, personalized, multi-media based, modern, tasteful, beautiful, creative, provocative. The remaining 14 items are appropriate for scaling the hedonic aspects of websites. Ranked from the largest p-value to the smallest, they are entertaining, humorous, full of action, exciting, surprising, playful, emotional, funny, motivating, challenging, intriguing, animated, inspiring, colorful.

For this set of items we apply several LR-tests by means of person-splits (twice 2-group random-splits, twice 3-group random split). The corresponding p-values range from 0.179 (median split) to 0.330 (2-group random split). Thus, they fit the PCM.

Item	**Location** ν_i	δ_{i0}	δ_{i1}	δ_{i2}	δ_{i3}
surprising	0.64280	-0.19135	0.91789	0.62664	1.21800
intriguing	-0.14716	-0.92523	-0.50483	0.00681	0.83460
inspiring	-0.17228	-1.08706	-0.09525	-0.23883	0.73202
playful	0.42209	-0.47539	0.51941	0.28856	1.35580
animated	0.39880	-0.15575	0.44310	0.14873	1.15912
funny	0.42476	-0.30505	0.36686	0.21649	1.42075
entertaining	-0.15026	-0.92691	0.26287	-0.55900	0.62201
motivating	-0.39728	-1.03638	-0.26578	-0.94980	0.66283
exciting	0.08451	-0.94097	0.24901	-0.16873	1.19874
emotional	0.63072	-0.30728	0.52412	0.92051	1.38553
colorful	0.39560	-0.81493	0.61036	0.33133	1.45564
full of action	0.72809	-0.16118	0.58907	0.90377	1.58071
humorous	0.20810	-0.39842	0.51877	-0.38957	1.10161
challenging	0.36545	-0.35770	0.52601	0.05496	1.23854

Table 1: Item Location and Category Intersection Parameters

Table 1 shows the item location parameters ν_i and the category intersection parameters δ_{ij} for the final item subset. These parameter sets allow for a detailed interpretation of each single item. A graphical representation including a histogram of the person parameters on the top is given in Figure 1.

The final items are heterogeneous in their locations, ranging from "motivating" ($\nu_i = -0.397$) on the left-hand site of the continuum up to "full of action" ($\nu_i = 0.728$) on the right-hand site. A website gets a high hedonism score if it is highly rated on as many items as possible. We note that it is irrelevant which items are chosen, since all of them comply with the sufficiency characteristic of the Rasch model and are appropriate for measuring hedonism. Compared to an intelligence test, that means that a highly intelligent person should not only be able to solve the difficult items, but also the easy ones. To illustrate this issue, we consider the possible - but rather unlikely - case that a website gets the highest score of four on the five lowest items (in terms of the location parameter) and a zero score on the remaining ones. This response pattern would sum to a score of four. Another site gets a score of four on the five highest items and zero on the remaining ones, which results in the same score. Since these raw scores are sufficient, both websites would get the same parameter and thus lie on the same position on Θ. Therefore, the Rasch model assures that it is eligible and fair to sum

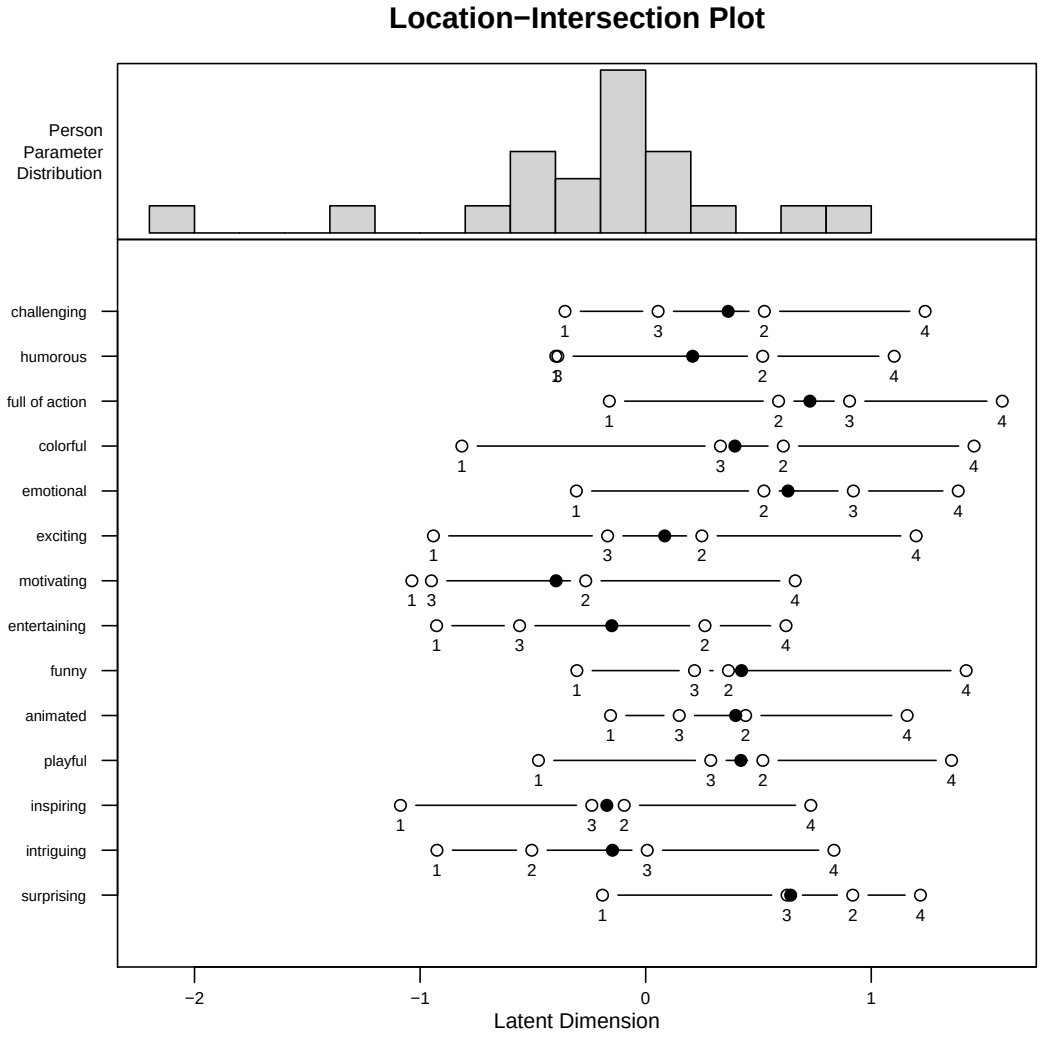


Figure 1: Plot of Location and Intersection Parameters

up the response scores of the final item subset.

As a consequence, location parameters "per se" are not a measure of quality. The heterogeneity in location parameters assures that we are able to map persons/websites with respect to their hedonistic affectivity on a wide range.

Location parameter allow for the interpretation of differences in affectivity according to the construct hedonism. For instance, the difference in item location between "emotional" and "funny" ($\Delta\nu = 0.620 - 0.424 = 0.186$) is approximately 2.25 as much than between "full of action" and "surprising" ($\Delta\nu = 0.728 - 0.643 = 0.0835$).

The category intersection parameters δ_{ij} as given in Table 1 denote the point on the latent continuum Θ at which the item category curves intersect. Figure 2 gives several examples of the underlying ICCs. When we look more closely at the item "emotional" we can see that category zero and category one intersect at a value of -0.307. That implies that as long as a website has an estimated hedonism score below -0.307, the probability of a zero score on this item will be higher than for any other category. As long as the level of hedonism is between $[-0.308; 0.524]$, the site will probably get a score of one on this item, and so on. Unlike the CTT, the researcher can interpret the results in a probabilistic manner. Therefore, IRT is sometimes referred to as *probabilistic test theory*.

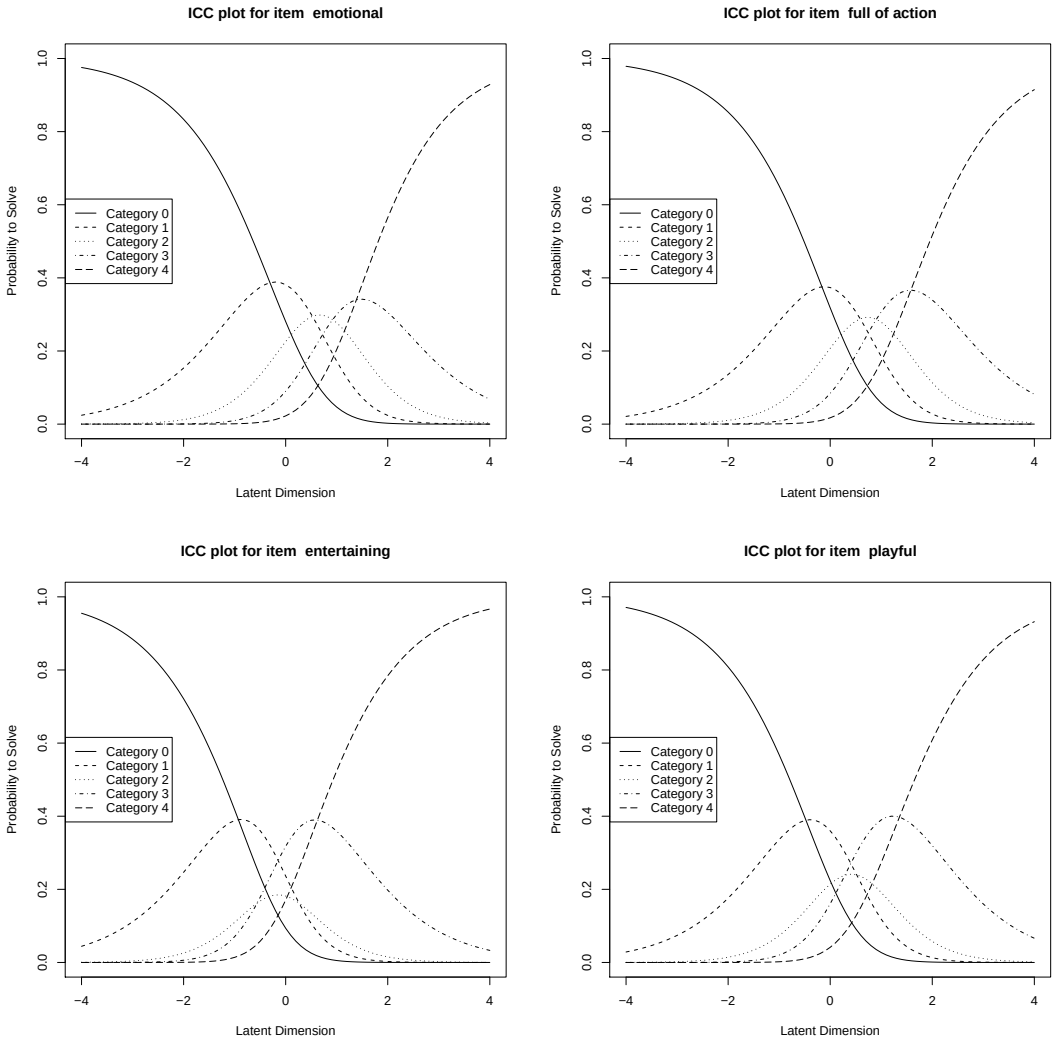


Figure 2: Item Category Curves

The items "emotional" and "full of action" possess a "regular" behavior in terms in increasing intersection parameters (as the category increases), i.e., $\delta_{i0} < \delta_{i1} < \delta_{i2} < \delta_{i3}$. This monotonicity property is not given for the two items at the bottom in Figure 2 or for a couple of other items. It is especially striking that $\delta_{i2} < \delta_{i1}$ for the items "entertaining" and "playful". That does not imply that there are not enough subjects with a score of three but shows that, conditional on the hedonism score, the probability for a response category three is lower throughout, compared to responses on other categories. This behavior is typical for "neutral" categories. In contrast, the more parsimonious RSM would not allow for non-monotonicity of intersection parameters. In addition, it restricts that the differences $\delta_{ij} - \delta_{ij'}$ are equal across all items. Thus, the ICCs are shifted horizontally according to the respective item location parameters.

The preceding analyses have illustrated which attributes are suitable for characterizing the hedonic aspects of websites, or, in other words, what might constitute the underlying rationale for Internet users to visit a certain site. As a side-effect, our analyses show the practical applicability of polytomous Rasch models in IS research. By concentrating mainly on CTT during the last decades, IS researchers have overlooked the potentials of IRT. By correctly applying this method in IS research, new insights can be gained about the content

domain of frequently used constructs.

5 Discussion and Conclusion

In this paper we present a probabilistic framework to measure latent constructs in IS by means of polytomous Rasch models. Even though these models were developed decades ago and are well-founded from a statistical perspective, they are not widely applied and researchers still rely on CTT. One reason for that might be that outside the psychometrics framework there is still a dearth of introductory books with applications. Until recently, another reason was the availabilty of user-friendly IRT software, since IRT models have not been implemented in standard software such as SPSS yet. As Borsboom (2006, p. 433) points out, some researchers are "monogamous" in terms of their software use, and so and there is little chance to convince them to use methods that are not clickable.

Many researchers who are outside of the statistics community have become more and more interested in open source platforms such as `R`, which provide a high degree of flexiblility. In turn, within the `R` community itself psychometricians have shown a growing interest in programming packages for IRT and related methods (see Mair and Hatzinger, 2007b).

If we look at historical developments in IRT from a methodological point of view, it is obvious that researchers have not made sufficient efforts to provide general frameworks of these models. Rather, the models have been developed somewhat separated from each other. Researchers have focused primarily on highly parameterized generalizations, which make the parameter interpretation even more difficult; or as de Leeuw (1998) points out: "A few of the generalizations seem to be motivated for the same type of reason Sir Edmund Hilary gave for climbing the Mount Everest: because it is there." Recently, de Boeck and Wilson (2004) embedded IRT models into the large framework of *generalized linear mixed models* (GLMM). This approach allows for the incorporation and interpretation of effects on both items and persons in a regression-type manner. Many IRT models, that were initially separated, now fit into this comprehensive framework.

A further promising approach for future IRT applications lies in the development and implementation of multidimensional IRT models, i.e. that items/persons can be mapped simultaneously onto more correlated dimensions (von Davier and Carstensen, 2007).

As we show in this paper, open source software packages allow for the successful usage of IRT in social science research, and especially in IS research. As soon as the conceptual understanding disseminates outside the psychometricians community and social science researchers learn how to correctly apply this method and how to interpret the results, these researchers have a powerful instrument for objective measurement at hand, one which overcomes several shortcomings of classical CTT.

References

Adams, R. J., Wu, M. L., and Carstensen, C. H. (2007). Application of multivariate Rasch models in international large-scale educational assessments. In von Davier, M. and Carstensen, C. H., editors, *Multivariate and mixture distribution Rasch models: Extensions and Applications*, pages 271–280. Springer, New York.

Agarwal, R. and Karahanna, E. (2000). Time flies when you're having fun: Cognitive absorption and beliefs about information technology usage. *MIS Quarterly*, 24(4):665–694.

Allport, C. D. and Kerler, I. W. (2003). A research note regarding the development of the consensus on appropriation scale. *Information Systems Research*, 14(4):356–359.

Alvarez, P., Lopez-Rodriguez, F., Canito, J. L., Moral, F. J., and Camacho, A. (2007). Development of a measure model for optimal planning of maintenance and improvement of roads. *Computers & Industrial Engineering*, 52(3):327–335.

Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika*, 38:123–140.

Andersen, E. B. (1995). Polytomous rasch models and their estimation. In Fischer, G. and Molenaar, I., editors, *Rasch Models: Foundations, Recent Developments, and Applications*, pages 271–292. Springer, New York.

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

Andrich, D. (1988). *Rasch models for measurement*. Sage, Newbury Park, CA.

Baker, F. B. and Kim, S. (2004). *Item response theory: Parameter estimation techniques*. Dekker, New York, 2nd edition.

Bock, R. D. and Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of an EM algorithm. *Psychometrika*, 46:443–459.

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika*, 71:425–440.

Carte, T. A. and Russell, C. J. (2003). In pursuit of moderation: Nine common errors and their solutions. *MIS Quarterly*, 27(3):479–501.

Cheung, W., Chang, M. K., and Lai, V. S. (2000). Prediction of Internet and World Wide Web usage at work: A test of an extended Triandis model. *Decision Support Systems*, 30(1):83–100.

Chin, W. W. and Marcolin, B. L. (2003). A partial least squares latent variable modeling approach for measuring interaction effects: Results from a monte carlo simulation study and an electronic-mail emotion/adoption study. *Information Systems Research*, 14(2):189–217.

Churchill, G. A. (1979). A paradigm for developing better measures of marketing constructs. *Journal of Marketing Research*, 16:64–73.

Cotte, J., Chowdhury, T. G., Rateshwar, S., and Ricci, L. M. (2006). Pleasure or utility? Time planning style and web usage behaviors. *Journal of Interactive Marketing*, 20(1):45–57.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16:297–334.

Csikszentmihalyi, M. (1990). *Flow: The psychology of the optimal experience*. Harper and Row, New York.

Davis, F. D., Bagozzi, R. P., and Warshaw, P. R. (1992). Extrinsic and intrinsic motivation to use computers in the workplace. *Journal of Applied Social Psychology*, 22(14):1111–1132.

de Boeck, P. and Wilson, M. (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. Springer, New York.

de Leeuw, J. (1998). Review of Fischer and Molenaar (eds.): Rasch models. *UCLA Statistics Preprint Series*, 202:1–4.

de Leeuw, J. and Verhelst, N. (1986). Maximum likelihood estimation in generalized Rasch models. *Journal of Educational Statistics*, 11:183–196.

Dekleva, S. and Drehmer, D. (1997). Measuring software engineering evolution: A Rasch calibration. *Information Systems Research*, 8(1):95–102.

Diamantopoulos, A. and Winklhofer, H. M. (2001). Index construction with formative indicators: An alternative to scale development. *Journal of Marketing Research*, 38:269–277.

Embretson, S. E. and Reise, S. (2008). *Item response theory for psychologists*. Lawrence Erlbaum, Mahwah, NJ, 2nd edition.

Fischer, G. H. (1974). *Einführung in die Theorie psychologischer Tests [Introduction to mental test theory]*. Huber, Bern.

Fischer, G. H. and Molenaar, I. W. (1995). *Rasch models: Foundations, recent developments and applications*. Springer, New York.

Füller, J., Bartl, M., Ernst, H., and Mühlbacher, H. (2006). Community based innovation: How to integrate members of virtual communities into new product development. *Electronic Commerce Research*, 6(57-73).

Haberman, S. J. (1977). Maximum likelihood estimates in exponential response models. *The Annals of Statistics*, 5:814–841.

Hambleton, R. K. and Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement*, 12:38–47.

Hirschman, E. C. and Holbrook, M. B. (1982). Hedonic consumption: Emerging concepts, methods and propositions. *Journal of Marketing*, 46(3):92–101.

Hoijtink, H. and Boomsma, A. (1995). On person parameter estimation in the dichotomous Rasch model. In Fischer, G. and Molenaar, I., editors, *Rasch Models: Foundations, Recent Developments, and Applications*, pages 53–68. Springer, New York.

Holbrook, M. B. and Hirschman, E. C. (1982). The experiental aspects of consumption: Consumer fantasies, feelings, and fun. *Journal of Consumer Research*, 9(2):132–140.

Huang, M.-H. (2005). Web performance scale. *Information & Management*, 42(6):841–852.

Lee, A. S. and Baskerville, R. L. (2003). Generalizing generalizability in information systems research. *Information Systems Research*, 14(3):221–243.

Lin, C. S., Wu, S., and Tsai, R. J. (2005). Integrating perceived playfulness into expectation-confirmation model for web portal context. *Information & Management*, 42(5):683–693.

Lord, F. M. and Novick, M. (1968). *Statistical theories of mental test scores*. Addison Wesley, Reading, MA.

Mair, P. and Hatzinger, R. (2007a). Extended Rasch modeling: The eRm package for the application of IRT models in R. *Journal of Statistical Software*, 20(9):1–20.

Mair, P. and Hatzinger, R. (2007b). Psychometrics task view. *R-News*, 7/3:38–40.

Mair, P. and Hatzinger, R. (2007c). CML based estimation of extended Rasch models with the eRm package in R. *Psychology Science*, 49:26–43.

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

Michell, J. (1999). *Measurement in psychology: A critical history of a methodological concept*. Cambridge University Press, Cambridge, MA.

Moon, J.-W. and Kim, Y.-G. (2001). Extending the tam for a World-Wide-Web context. *Information & Management*, 38(4):217–230.

OECD (2007). *PISA - The OECD Programme for International Student Assessment*. Organization for Economic Co-Operation and Development.

Petter, S., Straub, D., and Rai, A. (2007). Specifying formative constructs in information systems research. *MIS Quarterly*, 31(4):623–656.

R Development Core Team (2007). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Danish Institute for Educational Research, Copenhagen.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In *Proceedings of the IV. Berkeley Symposium on Mathematical Statistics and Probability, Vol. IV*, pages 321–333. University of California Press, Berkeley.

Salisbury, D. W., Chin, W. W., Gopal, A., and Newsted, P. R. (2002). Research report: Better theory through measurement - developing a scale to capture consensus on appropriation. *Information Systems Research*, 13(1):91–103.

Salzberger, T. (1999). How the Rasch model may shift our perspective of measurement in marketing research. *Marketing in the Third Millenium: Proceedings of the 1999 Australia and New Zealand Marketing Academy Conference (ANZMAC)*.

Salzberger, T. (2007). *Scientific measurement of latent variables in marketing research: An alternative framework*. Postdoctoral Lecture Qualification, Vienna University of Economics and BA.

Salzberger, T. and Sinkovics, R. R. (2004). Reconsidering the problem of data equivalence in international marketing research. *International Marketing Review*, 23:390–417.

Smith, R. M. (2004). Fit analysis in latent trait measurement models. In Smith, E. S. and Smith, R. M., editors, *Introduction to Rasch Measurement*, pages 73–92. JAM Press, Maple Grove, MN.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, 103:667–680.

Straub, D. W. (1989). Validating instruments in MIS research. *MIS Quarterly*, 13(2):147–169.

Thurstone, L. L. (1925). A method of scaling psychological and educational tests. *Journal of Educational Psychology*, 16:433–451.

Triandis, H. C. (1977). *Interpersonal Behavior*. Brooks, Cole, Monterey, CA.

Triandis, H. C. (1980). Values, attitudes, and interpersonal behavior. In Howe, H. E. and Page, M., editors, *Nebraska Symposium on Motivation 1979*, pages 195–259. University of Nebraska Press, Lincoln.

Tuerlinckx, F. and Wang, W. (2004). Models for polytomous data. In de Boeck, P. and Wilson, M., editors, *Explanatory item response models: A generalized linear and nonlinear approach*, pages 75–110. Springer, New York.

van der Heijden, H. (2004). User acceptance of hedonic information systems. *MIS Quarterly*, 28(4):695–704.

van der Linden, W. J. and Hambleton, R. K. (1997). *Handbook of modern item response theory*. Springer, New York.

Venkatraman, M. P. and MacInnis, D. J. (1985). The epistemic and sensory exploratory behaviors of hedonic and cognitive consumers. *Advances in Consumer Research*, 12(1):102–107.

von Davier, M. and Carstensen, C. (2007). *Multivariate and mixture distribution Rasch models: Extensions and applications*. Springer, New York.

Weber, R. (2003). Theoretically speaking. *MIS Quarterly*, 27(3):iii–xii.

Webster, J. and Martocchio, J. J. (1992). Microcomputer playfulness: Development of a measure with workplace implications. *MIS Quarterly*, 16:201–226.

Weiss, D. J. and Davison, M. L. (1981). Test theory and methods. *Annual Review of Psychology*, 32:629–658.