Thomas Rusch and Paul Benjamin Lowry and Patrick Mair and Horst Treiblmaier

Breaking Free from the Limitations of Classical Test Theory: Developing and Measuring Information Systems Scales Using Item Response Theory

Breaking Free from the Limitations of Classical Test Theory: Developing and Measuring Information Systems Scales Using Item Response Theory

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 latent constructs. The vast majority of IS studies uses classical test theory (CTT), but this approach suffers from three major theoretical shortcomings: (1) it assumes a linear relationship between the latent variable and observed scores, which rarely represents the empirical reality of behavioral constructs; (2) the true score can either not be estimated directly or only by making assumptions that are difficult to be met; and (3) parameters such as reliability, discrimination, location, or factor loadings depend on the sample being used. To address these issues, we present item response theory (IRT) as a collection of viable alternatives for measuring continuous latent variables by means of categorical indicators (i.e., measurement variables). IRT offers several advantages: (1) it assumes nonlinear relationships; (2) it allows more appropriate estimation of the true score; (3) it can estimate item parameters independently of the sample being used; (4) it allows the researcher to select items that are in accordance with a desired model; and (5) it applies and generalizes concepts such as reliability and internal consistency, and thus allows researchers to derive more information about the measurement process. We use a CTT approach as well as Rasch models (a special class of IRT models) to demonstrate how a scale for measuring hedonic aspects of websites is developed under both approaches. The results illustrate how IRT can be successfully applied in IS research and provide better scale results than CTT. We conclude by explaining the most appropriate circumstances for applying IRT, as well as the limitations of IRT.

KEYWORDS

Item Response Theory, Classical Test Theory, Scale Development, Rasch Model, Measurement, Measures, Hedonism, Reliability, Hedonic IS

1. INTRODUCTION

Social science research and information systems (IS) research produce a wealth of empirical papers that use survey or experimental data either to create new measurement scales or to apply previously validated scales to measure constructs. In most cases, the authors rely on fundamental measurement principles that have been developed and refined in classical test theory (CTT) over decades. Although several shortcomings of this approach are increasingly understood, the underlying measurement paradigm of CTT remains largely unquestioned in IS. In line with a recent call in IS literature to improve the methodological foundation of our domain [13]—the measurement and validation procedures [50]—in this paper, we present an alternative to CTT that opens up new perspectives for empirical IS research.

Psychometricians such as Spearman [81], [82], Thurstone [85], [86], Rasch [64], and Birnbaum [9] have formulated different statistical models to achieve the measurement of latent traits. Usually, *latent traits* pertain to any type of construct that cannot be directly observed. Two main approaches for measuring continuous latent traits emerged: CTT [e.g., 33, 45] and Factor Analysis (FA) [e.g., 95] on the one hand and Item Response Theory (IRT) [e.g., 44] on the other, with the former gaining widespread popularity.

Today, most research papers utilizing IRT can be found in psychology and educational testing, and at the same time the IRT paradigm is slowly but steadily gaining traction in social science and marketing research [73]. Several publications have clearly shown the advantages of this measurement approach [e.g., 28, 29], and thus have sparked new interest in using IRT in behavioral research [22, 27, 72, 75]. Despite these promising developments, so far, IS research has virtually ignored IRT, which might be due to the fact that IRT is frequently only associated with psychological testing. However, as Edelen and Reeve [20] have shown in their comprehensive study, “when used appropriately, IRT can be a powerful tool for questionnaire development, evaluation, and refinement, resulting in precise, valid, and relatively brief instruments that minimize response burden” (p. 5).

A few key example studies show that IRT and Rasch Models, which are often perceived as being

restricted to specific kinds of psychological testing, are in fact very versatile measurement methods that are applicable in a wide variety of disciplines. Rasch Models are a special class of IRT models that focus on the requirements for fundamental measurement and are relatively easy to understand; whereas, IRT in general deals with fitting flexible models to observed data.

An example in IS research includes a paper published in *Information System Research* in which they strived to understand software development practices [17]. The authors conclude that "The Rasch model analysis describes the likelihood of a practice deployment for any level of evolution and provides precise and meaningful measures" (p. 95). A marketing paper proposed a ten-item instrument for measuring customer satisfaction, which is a construct also frequently used in IS research [67]. Related examples from Marketing include brand equity [98] or the presence of gender item bias [74]. A Finance study used IRT to measure corporate social responsibility [57].

Moreover, Reise and Revicki [68] present several useful applications of IRT, including the assessment of data quality and the generation of item banks for hospital patients' questionnaires, which bears important implications for researchers interested in the healthcare industry. Another interesting example from the healthcare sector is given by Melas et al. [54] who illustrate with the help of IRT that the previously assumed poor correlation between attitudes toward evidence-based practice and communication technology is a methodological artifact rather than a substantive fact. Additionally, the current PISA study (Programme for International Student Assessment), which is conducted in most 60 OECD member countries (OECD, 2014), has successfully applied an extended version of the Rasch model [1]. Finally, Alvarez et al. [4] illustrate the versatily of the Rasch model in their publication on optimal road planning where they use it to obtain an objective measure of road conditions.

In this paper, we therefore explain why IS researchers should consider adding IRT to their existing pool of methods. Typically, when researchers measure latent variable(s), they strive to find a "good" set of items that allows for reliable, highly informative, and possibly invariant measurement of the underlying construct. Such measurement cannot be sufficiently guaranteed by CTT and related

approaches. The many models of IRT were developed to overcome this problem; to meet different goals and to allow different insight into the measurement process. The models' nature range from exploratory to confirmatory, from flexible to strict, from parametric to nonparametric (for an overview see [93]), and they try to meet different objectives in terms of what constitutes good measurement.

Objective measurement means "the repetition of a unit amount that maintains its size, within an allowable range of error, no matter which instrument, intended to measure the variable of interest, is used and no matter who or what relevant person or thing is measured" [66]. In this paper, we adopt and demonstrate the unique perspective of objective measurement typical for a class of IRT models, the family of Rasch models. Although Rasch models are restrictive in terms of item selection and model fit, they can provide a number of properties that are advantageous for scale development and substantive research based on these scales.

We argue that in the IS field certain conventions (such as treating measurement variables as metric) as well as the nature of CTT can be problematic in not meeting research goals because of the following limitations of CTT: (1) it assumes a linear relationship between the latent variable and observed scores; (2) the true score can either not be estimated directly or only by making strong assumptions; and (3) parameters such as reliability, discrimination, location, or factor loadings depend on the sample being used.

These limitations have a number of implications when used with categorical measures in behavioral IS research. For example, by assuming linear relationships, CTT treats a scale that is discrete and restricted, to say 5 values, as if it was stretching continuously from minus infinity to plus infinity. But visualizations of data derived from categorical measures show a very different behavior, for example, accumulation at certain values, gaps between values or more than a single peak. For these scales, the continuous assumption may only serve as an approximation. Another implication is that the sample dependence of parameters makes it hard to generalize results to a population, particularly if non-probabilistic sampling was used. Constant replication and revalidation of results derived from such

measures is needed to gauge their validity. Also, inference about the behavior of the units in question, about possible group differences, or the influence of a unit's characteristics can be associated with considerable bias.

In contrast, IRT offers five benefits, in that it: (1) allows nonlinear relationships; (2) allows appropriate estimation of the true score; (3) can estimate item parameters independently of the sample being used; (4) allows the researcher to select items that are in accordance with a desired model; and (5) applies and generalizes concepts such as reliability and internal consistency, and thus allows researchers derive more information about the measurement process.

As a demonstration of the applicability of Rasch models to IS research, we developed a scale for measuring hedonic IS, an area of IS research that has increasingly gained importance in recent years [18, 43, 92, 97]. For the purpose of this research, we initially create an item base that is as broad as possible to reflect the hedonic attributes of websites. To demonstrate the advantages that IRT models can offer, we perform an empirical comparative analysis of the scale results from a CTT versus Rasch perspective. Our goal is to find those items that measure hedonism as a latent construct unidimensionally and objectively, and to investigate how the underlying construct is measured by the items. Before demonstrating the empirical advantages of IRT scales and our example hedonic measure, we first provide the requisite background on CTT and IRT.

2. THE CONCEPTS AND ASSUMPTIONS OF CTT AND IRT

Conceptually, CTT and IRT strive to achieve the same thing—namely, inference about a continuous latent trait based on a number of manifest indicators (i.e., measurement variables). Both approaches are concerned with how to approach reliability, internal consistency, and the construct validity of scales; how to infer estimates of the latent trait value for each subject; and how to gain information about and assert certain properties of the measurement process. They mainly differ in the response model that is used for conducting inference: CTT uses techniques based on correlation, linear models, and multivariate normally distributed variables; whereas IRT approaches employ models for categorical

responses, use categorical association, and are concerned with multivariate discrete distributions. Thus, IRT models can be thought of as types of categorical factor analysis [cf. 7].

Therefore, the arguments pertaining to their difference or appropriateness in a measurement context are inherently statistical. Beyond that, both approaches offer similar insights or suffer the similar problems, specifically with regards to forms of validity other than construct validity. Even though IRT models were developed for categorical items, they are conceptually largely equivalent to approaches prevalent in IS.

In the remainder of this section, we first address the controversy of whether scales used in behavioral IS research (particularly Likert-type scales) are categorical or metric in nature. We then lay out the key differences between CTT and IRT for these measures, along with their strengths and weaknesses.

2.1 Which Scales are Inherently Categorical or Metric?

Before discussing CTT, it is critical that we address the issue as to what we believe constitutes categorical and metric measures. We base our arguments on a substantial base of theory and measurement articles, to which we refer the reader [3, 15, 19, 31, 40, 48, 49, 56, 83, 84, 90, 91]. Their key aspect is that categorical variables can have at least two different assigned values, that the same assigned values means things are the same, and that different assigned values means that things are different. How different and in what way may or may not be defined exactly. For example, it may be that three values stand at equal footing next to each other and we just know they are different (sometimes coined *nominal*), or it may be that there is some inherent ordering (coined *ordinal*) or that there is even transitivity (coined *strictly ordinal*). The more we know about what the differences between the assigned values mean, the more information we have.

In the case of metric variables we know a lot about the differences between the assigned numerical values. They are the result of an act of physical measuring (e.g., time, counts, length) and the assigned values relate to a clearly defined physical relation in reality: one second is twice as short as two seconds, and a debt of 100 USD is more debt than a debt of 50 USD by exactly the amount a profit of 51

USD is more than a profit of 1 USD. The key aspect of a metric variable is that differences between assigned values are constant in their meaning with respect to the underlying real relation the variable tries to capture. In other words, the size of the difference between any two assigned values of a variable has a meaning and is itself metric, so any re-scaling of the values will in the same fashion re-scale the difference. For example, it makes no difference regarding the real relationship between lengths whether we measure length in meter or centimeter. As one centimeter is 1/100 meter, the difference between one and two meters is 100 times the difference between one and two centimeters. Moreover, a metric variable also subsumes strict ordering and categorical uniqueness/exclusiveness for the assigned values as laid out before, hence it has a natural ordering (is therefore ordinal) and is unique and exclusive in the meaning of its values (we know 61 is not the same as 60, but 60 is). In summary, we view metric variables as a special, highly informative case of categorical variables. This helps us in motivating the IRT approach later on because while every metric variable must be inherently categorical, not every categorical variable is also metric.

When measuring attitudes or other latent variables in the social sciences, researchers often use measurement items or indicators which produce variables for which we simply do not know whether the differences between the assigned values fulfill the conditions laid out above. It is therefore a leap of faith to assume that a constant relation of the assigned numerals between two values of such an indicator corresponds to a constant relation in the real latent variable, or in other words, to assume they are metric.

A popular type of such indicators that are widely abused are the so-called Likert-type items [40]. Following our definition above, these items are always categorical and may be metric, *but they need not be*. Clearly, this subtle issue is the subject of a long disagreement, as many social science researchers—and virtually all IS researchers—treat Likert-type scales as metric scales for statistical convenience. This might work reasonably well but can also be completely wrong. It all boils down to whether it is safe to assume that the conditions for the indicator being metric hold.

This concern appears even when going back to the foundation of Likert-type scales, at which time

Likert himself made it clear that a proper Likert-type scale emerges from collective responses to a set of items, not the format in which responses are scored along a range [42]. Importantly, for the scale to be proper, Likert scaling *assumes* equal distances between the assigned numerals, and "all items are assumed to be replications of each other or in other words items are considered to be parallel instruments" [19, 31, 91, p. 197]. Whether this actually is the case cannot be inferred prior to data collection and is usually not checked afterwards. If it is, however, the results are often disheartening [for actual examples see, 15, 36, 88]. The correspondence of equidistance of assigned numerals for such a measure to equidistance of the underlying phenomenon becomes increasingly unlikely when considering that the items of a Likert scale are typically assigned progressive, but completely arbitrary numbers, and stand for numerical integers in which the progression represents "better". The fact that the distance between the integers is often chosen to be equal should not obscure the fact that this has nothing to do with whether the underlying relations are in reality also equidistant. If it were, then the latent continuum would have to be cut into regions of equal length by the indicator's values. Whether this applies cannot be asserted just by using a Likert-type scale.

Given the preceding, an imperative claim that we repeat from statistical literature is that until it is established that a certain scale (e.g., a Likert or any other rating scale) indeed fulfills the requirements for being metric for every data set, we cannot be sure that it does and thus should treat that variable as inherently categorical [15, 19, 48, 49, 56]. Hence, by treating such indicators as categorical, we are taking a safer stance: In the worst case, we might give up some information if the variable was indeed metric, which results in a (usually small) loss of statistical efficiency or power. Conversely, the worst case for treating a non-metric indicator as if it were metric is to make completely invalid conclusions, which is a far riskier stance taken by most IS research.

2.2 Classical Test Theory (CTT)

The common approach in scale construction is CTT [45] with its basic mathematical model being $X = T + E$ (1) where X denotes the observed overall score, T denotes the true overall score or the latent

construct, and E denotes the measurement error. Hence, the observed score is assumed to be a linear function of the underlying true score. This is a restrictive assumption that cannot be tested [23]. The idea of a linear relationship between an individual item response and the latent variable can further be generalized to factor models for one or more latent variables with different loadings or regression slopes per item, as seen in a common factor model or confirmatory factor model [e.g., 7]. The key issue in CTT and its generalizations is that the observed scores are linearly regressed on the latent constructs. A major problem with the assumption of a linear relationship is that if the latent trait is assumed to be on an interval scale, researchers treat the observed scores or sum scores as if they were interval scaled as well. This is not necessarily the case if the questionnaire uses ordinal indicators such as Likert-type scales.

2.3 A Critical Examination of Classical Test Theory (CTT)

A major critique of CTT is that the right-hand side of equation (1) is completely unknown; thus, to meet the equation, T and E can be chosen arbitrarily. Consequently, this equation is a tautology rather than a statistical model [23]. Regardless, researchers typically use reliability coefficients based on this basic expression. Reliability is defined as $\rho^2(X, T) = \sigma^2(T)/\sigma^2(X)$ [33]. Because T is unknown, we cannot compute its variance $\sigma^2(T)$. Consequently, additional assumptions are needed in terms of the measurement equivalence of test splitting. More often, reliability is commonly estimated by means of Cronbach's α [14], but this is actually an extremely limited reliability measure that is widely misused [78]. Notably, Cronbach's α indirectly includes the correlations between the items, and therefore is inherently a measure of the linear relationship between items. In the case of α and other CTT correlation-based reliability measures, the scores must be on an interval scale; otherwise, any correlation-based reliability will not be invariant.

The linearity and interval scale assumption is central to CTT. For example, when constructing a questionnaire or test, the whole process of item selection is based on correlation coefficients or transformations thereof. 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 nearly identical, such as 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. If factor models are employed, their estimates are also related to correlations.

Given this background, this section summarizes three main shortcomings of CTT for measurement variables, as addressed mainly in psychology [e.g., 10, 23, 35, 99] and marketing research [e.g., 71-73, 75]. First, the assumption of a linear relationship between the latent and observed scores is restrictive and is known not to necessarily represent empirical reality when it comes to psychological constructs [see, 24, 44]. Also, assuming such a linear function with different item locations implies that for certain values of the latent trait, no score is defined unless the item is metric and ranges from minus infinity to plus infinity. This is undesirable, because it restricts the span of the latent variable if categorical variables are used. If such a linear relationship is assumed, it is not congruent with the idea of the different locations of items. The same problem arises if different item discrimination (i.e., different slopes of the linear function) is allowed, as seen in FA. If one postulates a linear relationship, all items must have the same discrimination and location. This is called the assumption of τ-equivalent measures [45].

Figure 1 illustrates these issues. The latent "true" score is shown on the abscissa and the observed score on the ordinate. The relationship of the latent trait and the expected observed score is shown for four people (P1, P2, P3, P4) and three items. The dotted lines represent a person's latent true score, and the black dots the expected observed value of that person for a particular item. The τ-equivalent measure is depicted as the dash-dot line. Let us assume that we want to measure a latent score by means of a 4-point Likert scale with values ranging from 0 to 3. Let us further assume this scale corresponds to a restricted area of the latent trait between values of 0 to 6. Additionally, let us assume that the mapping of the linear trait values onto the Likert scale happens in such a way that the expected score is somehow discretized to obtain the observed score. Figure 1 shows the linear relationship between the observed and latent values. The expected observed values (i.e., solid lines) are depicted as functions of the latent trait values of three items, which follows from the assumption of a linear relationship. Items 1 and 2 have the same discrimination, but different locations from that of item 3, indicating that a higher value of the latent

trait is needed to score 3 at item 2 than at item 1. Figure 1 clearly illustrates the problem with CTT's

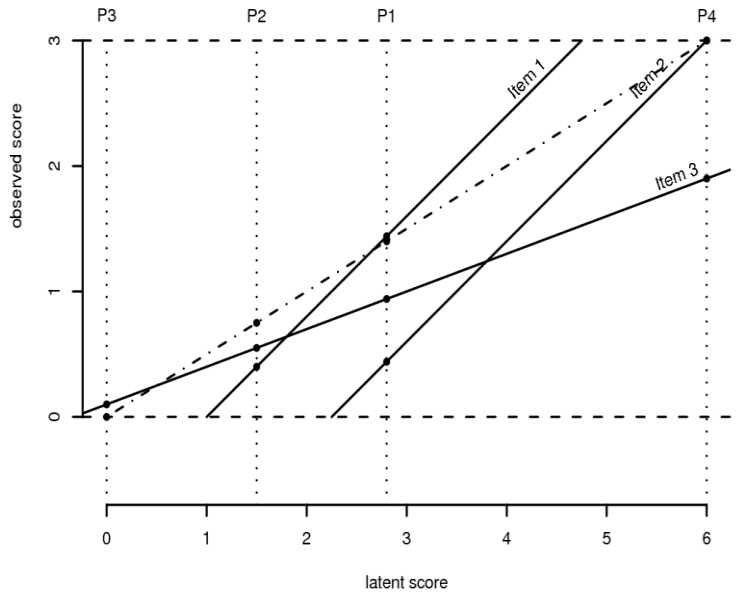

Figure 1. Representation of a CTT Model

assumption of a linear relationship. For example, P1, with a latent trait value of 2.8, has a positive expected observed score for all three items of the scale. This is not true for P2. For this person with a latent trait value of 1.5 we expect a negative observed score, which is impossible to achieve. This occurs because the item has a positive expected score in a different area on the latent trait. To apply a score to this latent value for item 2 means that one had to take zero, which leads to an error that will always be positive and, consequently, the expected observed score will not be the true score anymore. The same applies for person P3 with the lowest of the latent trait values, with respect to items 1 and 2. Item 3 is the only item that allows positive expected observed scores for all the people in the example, but at the same time has the lowest discrimination parameter and might therefore be considered the "worst" item in this scale.

Additionally, item 3 also measures latent trait values beyond our restricted area, which means that if we stay restricted, it is not possible to achieve an expected observed score of 3, as the expected score of P4 (i.e., the person with the highest latent trait position) illustrates. This item maps the restricted latent

trait onto a 3-level Likert scale only. These problems will always appear if any two items have different slopes or locations, which is the assumption of congeneric measurement in CTT. The only plausible way to construct items that have a linear relationship with the restricted latent trait is to have items with equal slopes and locations, such as the item with the dashed-dotted line. This is the τ-equivalent measure. Only this item allows for every restricted latent value to be assessed with this 4-point Likert scale. Consequently, a scale with CTT assumptions must consist of items that are fully interchangeable. Although more items will then increase the reliability of the test, no additional information regarding the person's latent value will be gained.

Second, in CTT the true score or factor score cannot be estimated directly, but only via additional assumptions regarding the item-specific true scores. A scoring rule (e.g., simple or weighted sum) is implicitly assumed to be correct, but its adequateness cannot be tested. Furthermore, the simple sums of the observed scores are often taken as an estimate of the person's latent trait value or the item's location. This approach equates the expected true values with the sum of the observed scores. However, it is possible that a person with a lower score in a test will have a higher position on the latent trait. This could be the case if this person fakes an answer, whereas the person with the higher location answers truthfully. Therefore, using the sum of observed scores is not necessarily appropriate for measuring this empirical reality.

Third in CTT, parameters such as reliability, discrimination, location, and factor loadings depend on the sample being used, which implies different reliabilities as well as different factor loadings of an item set for both homogeneous and heterogeneous samples. Hence, it is frequently the case that different numbers of factors for different samples emerge. They apply only to the sample at hand and are unbiased for the population of interest only if the sample is a true random sample and representative for the population of interest [e.g., 21, 23]. If we want to estimate the location of a person on a latent trait, that value depends on the sample of items used for measurement and on the other people who are being assessed. Depending on the reference population, a person will also have a different position on the latent

trait, even if the random sample is representative. Thus, such measurement can never be invariant, let alone objective. This poses a problem when different groups are compared by summary statistics that depend on the sample.

These shortcomings are inherent in CTT and cannot be resolved without adopting a different measurement paradigm. However, CTT is not an incorrect approaches per se. Instead, it is the method of choice when working with metric scales; this choice is also true for structural equation modeling (SEM) and confirmatory factor analysis (CFA), which are models that often need additional distributional assumptions. In short, CTT provides a rich framework for conducting analyses if two key assumptions hold: (1) it is theoretically/empirically justifiable that the observed scores lie on a metric scale, and (2) the functional relationship is linear. Moreover, with Likert-type scales, the measurement is better treated as categorical than treated as interval, which contrasts with extant IS research practice.

2.4 Item Response Theory (IRT)

To illustrate the underlying rationale of IRT models, let us assume the simplest case of items with two categories (i.e., a *dichotomous* item), coded with "1" and "0." Let β_i, a parameter connected with an item (i.e., item parameter) denote the location of an item *i* (*i*=1,...,K) on the latent trait such that the higher its value the less the probability of scoring 1. More accurately, β_i is the value on the latent trait where scoring 1 has a probability of 0.5 for this item. Let θ_v denote the position of person v (v=1,..,N) on the latent trait. If $\beta_i = \theta_v$, the probability of scoring 1 is 0.5 for that person. We therefore get a (0,1) people $\times$ items data matrix X of dimension N $\times$ K. The item response patterns $X_i$and person response patterns X_v are indicators for β_i and θ_v. Other than in CFA, no distributional assumptions of latent traits need to be imposed, but they can be. Given that the items are ordinal, the patterns X_i and X_v are still on an ordinal level, but β_i and θ_v lie on a metric scale.

More generally, let B denote a matrix S x K of S different item parameters in columns (e.g., discrimination, location) and the respective values of these parameters for each of the K items in rows. Let Θ denote the matrix that gives the positions on the latent dimensions that underlie the behavior in a

certain situation. Each column refers to a specific latent dimension and its entries to the location of people on that latent dimension if presented with all the K items. The basic functional relationship is then $P(X = x) = f(B, \Theta)$ with *f* being the item response function (IRF) or item characteristic curve (ICC). Different IRT approaches exist in terms of the number of item-related parameters or person-related parameters as well as the functional relationship. For instance, in addition to the location parameters β_i, the researcher might wish to allow for item-discrimination parameters α_i or guessing/faking parameters γ_i, which in some situations might be more realistic. For both item- and person-related parameters, it is the case that if a multidimensional construct is used, every person could have a different latent trait value on all dimensions, and every item might measure each trait to a certain degree. The traits can also be correlated. The functional relationship between the probability of scoring in a certain category and the person's position on the latent trait is usually allowed to be nonlinear and can be pre-specified (e.g., via a logistic function) but also estimated (e.g., via kernel smoothing).

Depending on the degree of parameterization and the overall goal of the analysis, two conceptually different approaches in IRT exist:

1. Item selection and confirmatory approach: In this approach, the aim is to find items for which a certain IRT appears to hold. These items may then exhibit various properties of these models, such as the "fairness" of comparing people, the sample independence of the estimates, different discrimination ability or heterogeneous locations. These models may allow for objective measurement [64].

2. Modeling and exploratory approach: If researchers are not primarily interested in selecting items in a very restrictive manner, but instead wish to analyze a person's response behavior and therefore the scale and its items, then they would take into account higher parameterized models like two-parameter logistic (2PL) or three-parameter logistic (3PL) models, multidimensional models, models with covariates [16], or nonparametric models [e.g., 63].

2.5 Advantages of IRT

As was noted earlier, when the goal is to measure a latent construct, CTT and related methods can lead to serious problems, particularly with range-restricted/categorical items. Borsboom [10] asserts 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" (p. 429). That is exactly what IRT does, and in doing so overcomes several limitations of CTT.

First, the linear relationship between the indicators and a categorical response, which is assumed in CTT, is usually not appropriate. Clearly, using transformations such as polynomials in CTT would be possible, but they require methodological and domain specific knowledge and only cover a fraction of possible relationships. Instead, in IRT, directly nonlinear relationships are used. Such a nonlinear function is more general and can subsume a linear relationship. In IRT, the nonlinear function that relates the probability of observing a certain response to an individual item with the latent trait is called the *item response function* (IRF) or *item characteristic curve* (ICC). This function enables flexible specifications of the theoretical relationship between the underlying trait and the items, given the response format (dichotomous or polytomous), contexts, or theoretical assumptions about the response process (e.g., dimensionality). Additionally, from an empirical point of view, the higher flexibility of IRT models allows for a close fit to be achieved between a function and the data. For example, if the real relationship is linear in the interval (0,1), it is possible to fit a near linear function with a 2PL model [9], whereas the opposite is not true. Furthermore, IRT models allow assessment of the adequateness by means of statistical goodness-of-fit tests and fit indices.

Second, IRT allows analyses to 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. IRT models usually allow for an estimate of the underlying latent trait value that incorporates all the available information from the data. They also enable researchers to define an appropriate scoring rule to adequately represent empirical relationships. In particular cases, they even permit the use of the sum

scores as the scoring rule, such as in Rasch measurement. In CTT, the weighted sum of the scores cannot generally be the appropriate scoring rule for the true score if dichotomous or polytomous items are used [6].

Third, if IRT models hold, they allow for consistent estimation of parameters irrespective of the sample composition used to estimate the parameters (item and person parameters alike). IRT allows researchers to assess whether a model can be assumed to hold and thus to derive the statistical properties that the model entails. For instance, if the model holds for the population, the parameters estimated from an infinite number of items have the same expected values in the population regardless of what items and sample has been used. Consequently, if an IRT model holds in a population, using only some (and possibly different) items to estimate people parameters is perfectly valid and leads, on average, to the same estimate as if other items were used. Any comparison of people will be asymptotically independent of the items being used and who else was in the sample.

Fourth, when selecting items to construct a scale, IRT enables the researcher to select items that are in accordance with a desired model. Out of a pool of possible items for the scale, the ones that conform to a Rasch model might be selected to ensure that its measurement properties apply. This selection is guided by the usage of statistical tests as well as graphical procedures. In CTT, items are often chosen based on sample dependent measures. It is possible that both approaches lead to similar or equivalent scales, but this need not be the case.

Fifth, concepts such as reliability and internal consistency are applicable to IRT models. The reliability and internal consistency of a set of items will be high if a one-dimensional IRT model holds, because all items measure the same trait; however, the reverse is not necessarily true. IRT models also allow researchers to gain more information regarding how an individual item measures and to investigate the suitability of an item over the distribution of the latent trait or for the respondents' latent trait values. Specifically, IRT allows the researcher to assess the standard error of estimation for every single item as a measure of precision. Therefore, confidence intervals for an estimate of each specific latent trait value can

be calculated. The standard errors and the resulting confidence intervals will differ between latent trait values or respondent locations. This is in accordance with the empirical finding that measurement in middle regions of a latent trait is more precise than in extreme regions [34]. Other than CTT, IRT does not assume this precision to be constant.

In IRT, it is further possible to calculate an item's information, which tells researchers how much knowledge about the areas on the latent trait they can derive from an individual item. This can be seen as a more general concept of precision as compared to reliability, because it really shows how much information a specific item carries for different latent trait values. For example, an easy item will not have much information about the latent trait area of a genius, but will have much information about people for whom solving easy items is challenging. It is thus known how well a scale can assess different peoples' locations on the latent trait. Precision is not constant, but can be different for different values of the latent trait. Similarly, it is possible to include items that measure the entire latent trait area of interest with high precision. That is, items can be heterogeneous in terms of which latent values they will measure. Moreover, an item that measures the same area of the latent trait as another item might have less information, and could therefore be excluded if there are length or other restrictions that ask for item removal. One can even assess the information of a whole scale, and that information can be compared to another scale measuring the same construct.

Sixth, with IRT it is possible to obtain detailed information on an item and person level simultaneously. Each item *i* and each person v is assigned one or more parameters (i.e., the location of item β_i and position on trait θ_v) that allow for a probabilistic analysis of the response behavior. Item and person parameters lie on a metric scale, which makes it possible to interpret the distances between items and people on Θ. This is especially noteworthy, since the observed responses are on an ordinal scale. Certain IRT models actually enhance the scale level. If the items and people are on the same scale, then statements about the response probability of person v on item *i* can be achieved. This means it is possible to predict the behavior of a person on a certain item.

2.6 Additional Properties of Rasch Models

Rasch stated that the objectivity of comparisons is a basic requirement [65], and he formulated the epistemological theory of specific objectivity (SO): objective because any comparison of a pair of parameters (items/people) should be independent of any other parameters or comparisons; specifically objective because the comparison made is relative to some specified frame of reference [5]. That is, under SO, two people v and w with abilities θ_v and θ_w are comparable independently from the remaining people in the sample and independently from the item subset in which they were 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 people in the sample [51]. To achieve this, very strict requirements are applied to these IRT models, and these restrictions lead to scales that exhibit extraordinary measurement qualities.

Rasch [64] presented a probabilistic model that can be used to study the response behavior of individuals on dichotomous items. It poses a logistic relationship between the ability θ_v of a person v and the probability of a correct response on item *i*. Each item gets a difficulty parameter β_i. The formal representation, which is known as the Rasch model, is

$$P(X_{vi}=1|\theta_v,\beta_i)=p_{vi}=\frac{\exp(\theta_v-\beta_i)}{1+\exp(\theta_v-\beta_i)}$$

where $P(X_{vi}=1)$ is the probability that person *v* answers 1 on item *i*. For simplicity, we will drop the indices in the following sections.

Figure 2 represents the basic ideas in terms of the ICC for two items. The probability of answering "1" is depicted on the ordinate: the abscissa displays the latent trait value. The probability of an observed score of 1 (solid line) and 0 (dotted line) for item *i* as a function of the latent trait, the ICC, is shown. For item j only the ICC for score 1 is depicted. The vertical solid lines represent the item location on the common scale of the item and latent trait (i.e., P(X=1)=0.5). For item j, a higher latent trait value is needed to achieve the same probability of observing score 1 than for item *i*. For item *i* with β=-0.3, both the probability of observing "0" (dotted line) and the probability of observing "1" (solid line) as a logistic

function of the underlying latent trait value is shown. These two lines intersect at P(X=1)=0.5, which is by definition the "location" β of item *i*. This value can also be interpreted as the threshold at which it becomes more likely to score 1 than 0. One can see that the higher the position of a person on the latent trait, the higher the probability of scoring 1 becomes, and vice versa. For item j, only the probability of scoring 1 is shown because P(X=0)=1-P(X=1). This item has a higher location (β_j=1) than item *i*, which means the probability of scoring 1 is lower than for item *i* for any given latent trait value. It is noteworthy that in this case, both items have the same discrimination; that is, they share the same "slope" of the logistic curve. Hence, Rasch models do not allow the logistic curves to cross.

Because of the functional relationship, no matter what the latent trait value is, a probability for a score is always defined. Additionally, we can assess the whole latent trait as long as we have enough items that are different in terms of their locations. Another interesting issue is that in the middle region—around the item's location—measurement is practically linear; but for the extreme regions, the item is not able to distinguish well between people. It can also be seen that people's abilities and item difficulties lie on the same scale. Furthermore, the intersection point cuts the latent trait into a region that corresponds to score 0 and score 1, which means that the model also maps dichotomous responses onto a metric scale.

Figure 2 also shows how restrictive the dichotomous Rasch model actually is. It requires that (a) there is only one underlying latent trait, (b) an equal discrimination parameter for different items exists, (c) the ICC is a logistic function, and (d) the probability of a certain score depends solely on item location and person location on the latent trait and, most important, (e) "local independence" which denotes the independence of each individual item response conditional on the person's latent trait value. The interdependency and statistical relationship between the observations of the different items is solely due to the position of the person's location on the latent trait. Therefore, concepts that are based on correlations between items, as in CTT, are seen as spurious. For example, a highly aggressive person will often agree to items that ask about violent behavior, whereas a pacifist will not. If the latent trait value is held constant, any relationship between the items disappears. As opposed to CTT, all of these restrictions

can be tested.

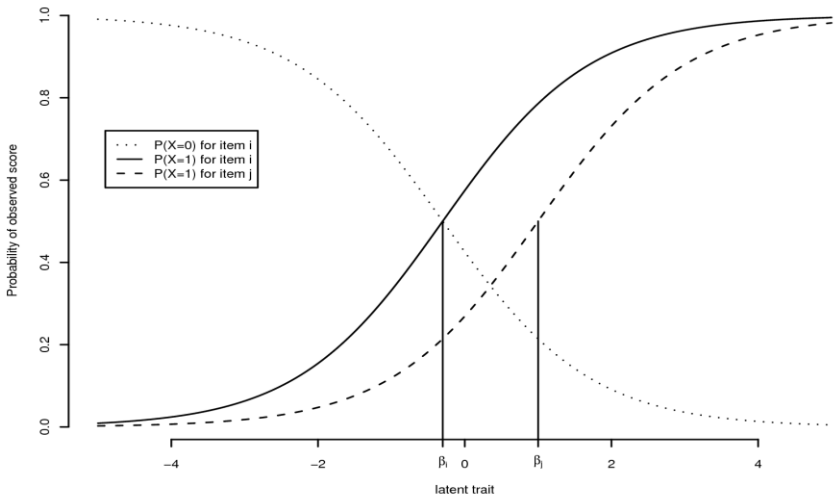

Figure 2. Representation of a Dichotomous Rasch Model for Two Items

Because of its restrictiveness, the Rasch model is not flexible enough for modeling purposes, but if all items conform to it, this model has some remarkable features, as previously described. To achieve model fit, one would eliminate items that contradict at least one of the Rasch model assumptions. Item selection can be conducted by different means, for example, by residual-based item fit statistics [80] or Wald tests [30]. Those items that remain in the final homogeneous item subset measure the latent construct in an objective manner.

In many practical situations, dichotomous item responses are too restrictive. This is especially true in social science research, where Likert-type scales are commonly used for assessing individuals' attributes. For such polytomous items, the model outlined above can be generalized. One popular extension is the partial credit model (PCM) [53], which we will focus on in the following sections.

Using PCM, all of the properties and assumptions of the dichotomous Rasch model still apply. Every ordinal item i with m_i as the number of categories is described by h-1 cumulative intersection parameters β_{ih}, which map the categories onto the latent trait. Thus, the PCM can be regarded as an adjacent-categories logit model [89]. For each category, there is a probability of scoring in each category

as a function of the latent trait. Basically, it estimates log-odds for a certain category h with respect to category h-1. An important issue, therefore, is the interpretation of the item-category parameters β_{ih}. These parameters are often transformed into category intersection parameters δ_{ij} with $j = 0 \ldots m_i$. If we estimate the PCM, the item categories are converted into intersection parameters as $\delta_{i0} = -\beta_{i0};\ \delta_{i1} = \beta_{i0} - \beta_{i1};\ \delta_{i2} = \beta_{i1} - \beta_{i2}$, etc. The parameters δ_{ij} refer to the points on the latent trait where the ICCs intersect. Based on these intersection parameters, we can compute item location parameters v_i in terms of $v_i = m_i^{-1} \sum_{j=0}^{m_i} \delta_{ij}$. Within the context of item selection to construct a scale, the main focus is on the item (-category) parameters that we can estimate independently from the person parameters if Rasch models are applied. In this case, we are not primarily interested in the estimation of θ. Our aim is to establish a homogeneous subset of items that allows for an objective measurement of a latent construct. To score people, a useful scale should have a wide range of items in terms of their locations.

3. DEMONSTRATION OF SCALE DEVELOPMENT IN AN IS CONTEXT

In this section, we demonstrate the applicability of the Rasch-type scale construction and measurement in IS by constructing a scale to measure hedonic IS. Hedonism, a powerful form of intrinsic motivation, has gained a lot of attention in the IS community, and several non-utilitarian constructs (i.e., non-extrinsic motivation) have been integrated into various theoretical models as its importance has become clearer. These constructs include perceived affective quality, cognitive absorption, perceived enjoyment, and perceived playfulness [2, 39, 46, 47, 77, 92, 94, 97, 101], and they frequently exhibit similar or identical items. Using a substantial number of multi-item scales leads to a vast amount of items with unclear measurement properties. Hedonism is therefore an ideal example to illustrate the strengths of IRT, which lie in the detection of the "measurement scope" and the suitability of the respective items. We therefore consider hedonism as the perfect context in which to illustrate the practical applicability of Rasch models. As Burton-Jones and Straub [11] suggest, we created our scale with a specific context in mind, which in our case is websites.

This scale is supposed to measure only one latent dimension. All of the scale items should be

able to assess it. To provide a more useful demonstration, we create two scales: one using a CTT approach, and one using PCM. The latter analysis will serve as a guideline for illustrating how scales can be constructed using an IRT approach.

3.1 Data Description

We used several steps to collect and clean the data. To ensure that the attributes represented all facets of the concept under investigation (i.e., content validity), we followed the instructions from Moore and Benbasat [55] and 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 being to come up with as many attributes as possible without any evaluation or rating. Subsequently, we used the same panel of experts to group the items they chose and to filter out synonyms, which resulted in a total of 26 items.

After performing 10 preliminary tests to ensure that the items were comprehensible, we conducted an online survey in which a convenience sample of 291 Internet users rated the importance of those attributes for measuring hedonic concepts. We used a 5-point scale with a range from zero ("not important") to four ("very important") to assess the significance of the single attributes. Therefore, the data matrix **X**, which we used for all subsequent analyses, consisted of 291 subjects and 26 items.

3.2 Descriptive Analysis

Table 1 shows the descriptive statistics of the 26 items for the 291 respondents of the sample.

3.3 CTT Analysis

We performed the computations in R [62], with the packages "psych" [69] for exploratory factor and reliability analysis, and "sem" [26] for confirmatory factor analysis. To calculate the polychoric correlations we used the package "polycor" [25]. In CTT, the first problem is to find out how many latent factors may be underlying our items. For our purpose, we aimed for one factor only. No clear-cut solution exists as there are various criteria to help choose the number of factors. To confirm the appropriateness of the data, we used the Kaiser-Meyer-Olkin Measure of Sampling Adequacy (0.88) and Bartlett's test of

Table 1. Descriptive Statistics

Name	Number	Mean	SD	Median	SE
Creative	1	2.68	1.21	3	0.07
Surprising	2	1.52	1.36	1	0.08
Intriguing	3	2.43	1.27	3	0.07
Inspiring	4	2.45	1.31	3	0.08
Playful	5	1.81	1.35	2	0.08
Animated	6	1.85	1.42	2	0.08
Multimedia based	7	2.45	1.28	3	0.07
Funny	8	1.84	1.37	2	0.08
Entertaining	9	2.45	1.37	3	0.08
Provocative	10	0.97	1.22	1	0.07
Motivating	11	2.75	1.23	3	0.07
Beautiful	12	2.47	1.21	3	0.07
Exciting	13	2.19	1.32	2	0.08
Frisky	14	1.00	1.24	1	0.07
Modern	15	2.75	1.26	3	0.07
Emotional	16	1.58	1.30	1	0.08
Colorful	17	1.82	1.30	2	0.08
Full of action	18	1.50	1.29	1	0.08
Humorous	19	2.08	1.43	2	0.08
Challenging	20	1.88	1.40	2	0.08
Interactive	21	2.53	1.18	3	0.07
Customized	22	2.29	1.28	2	0.07
Personalized	23	2.15	1.28	2	0.08
Tasteful	24	2.86	1.12	3	0.07
Plain	25	1.95	1.35	2	0.08
Suitable for children	26	1.77	1.49	2	0.09

SD ... Standard Deviation SE ... Standard Error, n = 291

sphericity ($p < .001$) [87]. We selected principal axis factoring of the polychoric correlation matrix and used the scree plot criterion as well as the very simple structure (VSS) criterion [70] to determine the number of factors.

There was a dramatic drop in explained variance after the first factor is extracted. Thus, the scree plot as well as the VSS criterion supported extracting only one factor. To verify this, we conducted a CFA

with one common factor. All items were allowed to load freely on the common factor. Using a significance level of 5%, we found that the loading of one item, "plain," was not significantly different from 0; thus, we deleted the item and refitted the CFA. The fit indices suggest that the model does not fit the data very well. The root mean square error of approximation (RMSEA) equals 0.111, which is above the recommended upper bound of .07 (some authors go as high as .1) and the comparative fit index (CFI) equals 0.725, which is well below the recommended lower threshold of 0.9 (some authors even recommend .95) [38]. Similarly, all other fit indices (e.g., Goodness-of-fit index, NFI, Tucker-Lewis NNFI, CFI, AIC) indicate a poor fit.

We also tried a two-factor solution where all of the items were allowed to load freely on one of the factors without cross loadings. Additionally, the two factors were allowed to correlate freely. The fit was better than that of the one-factor CFA (RMSEA=0.102, CFI=0.768) but it was not good enough to confirm a two-dimensional structure. Because our aim was to derive a single scale for measuring hedonic websites, we proceeded with deleting items that had small loadings on the factor from the one-factor solution, until the fit became worse again. The best fit was achieved after the deletion of the items "multimedia based," "beautiful," "modern," "interactive," "customized," "personal," "tasteful," and "plain" which left 18 items. Even this best FA model does not fit well to the data (RMSEA=0.106, CFI=0.835). Nevertheless, the selected items are suitable for the one-dimensional measurement of hedonic information systems within a CTT framework. Our scale has an impressive Cronbach's α of 0.91 and all loadings are significant. Appendix A shows the one-factor solution before and after item selection and the two-factor solution.

3.4 IRT Analysis

We performed all computations with the eRm package [51, 52] in R, which uses CML estimation. To achieve a final set of items, we used the following steps:

1. Estimate PCM item and person parameters.
2. Compute item-fit statistics based on residuals.

3. Eliminate the item with the least fit (i.e., highest item-fit statistic).
4. Compute LR test for different person sub splits. If LR is significant, go back to step (1) and eliminate items. Otherwise, the procedure stops and the final model is obtained.

When the data did not fit the PCM, we eliminated items successively and re-fitted the model. It is a peculiarity of the item selection approach that data are actually fitted to a model, not the other way round. Other IRT models allow for the conventional statistical approach of model fitting, but we decided to use a Rasch model, which is comparatively easy to understand and ideally suited for the task at hand. The result was a set of homogeneous items that comply with the restrictive Rasch criteria. This means that they all measure the same latent trait (i.e., one-dimensional), that the sum of the scores is the appropriate measure of the underlying latent trait (both for items and people) and that the estimated parameters are sample-independent (i.e., specific objectivity holds) if the model holds in the population. The reason for fitting the LR test after each step is that this statistic, which is a global model test, evaluates the model fit of the whole item set. 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 suitable for assessing model fit.

We started our analysis with the same total set of 26 items that we used in the previous section. We eliminated, based on the procedure described above, the following items in this order: “plain,” “suitable for children,” “interactive,” “customized,” “personalized,” “multi-media based,” “modern,” “tasteful,” “beautiful,” “creative,” “provocative,” “inspiring,” “intriguing,” “colorful,” and “animated.” The remaining 11 items are appropriate for scaling the hedonic aspects of websites within a Rasch framework. Ranked from the smallest to the largest item fit statistics, they are “frisky,” “humorous,” “entertaining,” “full of action,” “exciting,” “surprising,” “emotional,” “playful,” “challenging,” “funny,” and “motivating.” For this set of items, we applied a small simulation of 40 LR tests by means of person-splits (2-group random-splits and 3-group random splits). Table 2 shows the item location parameters v_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. For visual inspection of the common latent scale, the location and category threshold estimates are displayed in the lower part of Figure 3, together with the estimated person parameter distribution.

Table 2. Location and Threshold Parameters of Items Selected in the Rasch Model Approach

Item	Location	Threshold1	Threshold2	Threshold3	Threshold4
Surprising	0.55017	−0.3765	0.8126	0.55318	1.21139
Playful	0.31256	−0.6884	0.39266	0.20719	1.3388
Funny	0.31414	−0.5191	0.23827	0.13385	1.40349
Entertaining	−0.3055	−1.1942	0.0799	−0.6724	0.56484
Motivating	−0.5825	−1.3400	−0.4936	−1.0895	0.5932
Exciting	−0.0518	−1.1906	0.08804	−0.2669	1.16215
Frisky	1.15526	0.46667	1.40648	0.66995	2.07795
Emotional	0.53599	−0.5029	0.41212	0.84531	1.38936
Full of action	0.64045	−0.3493	0.48214	0.8303	1.5987
Humorous	0.08177	−0.6286	0.3733	−0.4814	1.06374
Challenging	0.25191	−0.5727	0.39598	−0.0284	1.21273

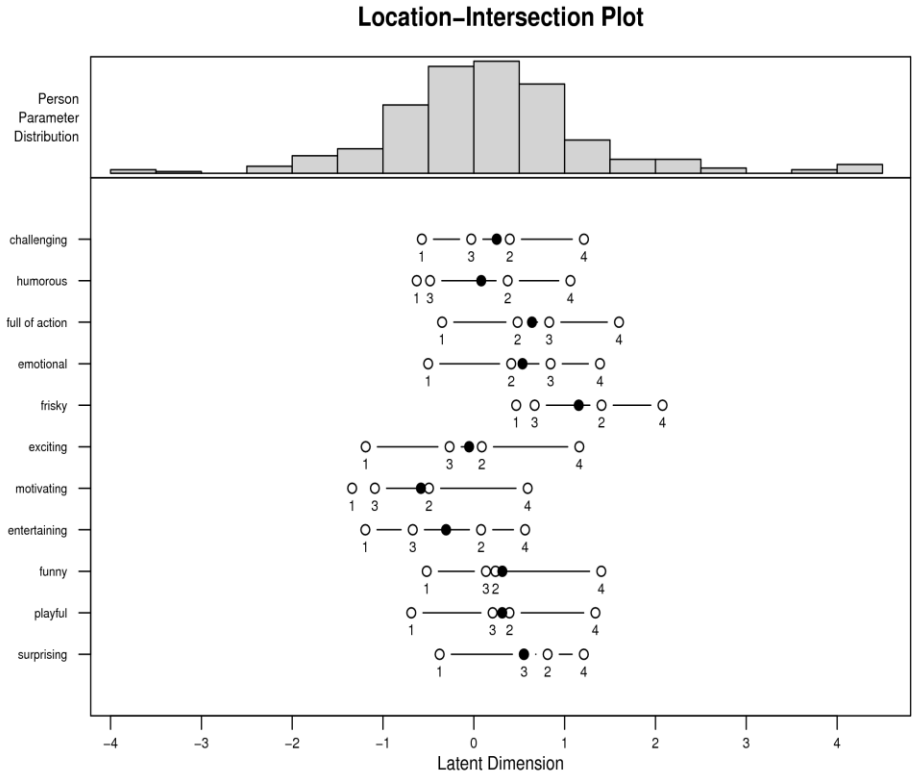

Figure 3. Location of the Item Categories on the Common Latent Trait Scale

The final items are heterogeneous in their locations on the "importance for measuring hedonism" latent scale —ranging from "motivating" (ν_i=-0.582) on the left-hand side of the continuum up to "frisky" (ν_i=1.155) on the right-hand side. We note that when measuring hedonism, it is irrelevant which items from this pool are chosen because all of them comply with the Rasch model, and thus are appropriate for measuring hedonism. Also, since raw scores are the appropriate scores for a Rasch-type model, all items/people with the same raw score would get the same parameter, and thus lie on the same position on Θ, the latent trait.

The higher the location of an estimated location parameter, the more important the item is considered to be for hedonism. Location parameters allow for the interpretation of differences in importance according to the construct hedonism. For instance, the difference in item location between "emotional" and "funny" (0.535-0.314=0.22) is approximately 2.4 times as much as between "full of action" and "surprising" (0.64-0.55=0.09). This means the latter are 2.4 times more similar in terms of the amount of the construct the items represent than the former.

The category intersection parameters δ_{ij} denote the points on the latent continuum Θ at which the category characteristic curves (CCC) intersect. Figure 4 shows several examples of the underlying CCCs. Each line visualizes the probability of observing a certain response as a function of the latent trait values. Examining the item "emotional," in the upper left of Figure 4 the categories 0 and 1 intersect at a value of -0.502. This implies that as long as a website has an estimated hedonism score below -0.502, the probability of a zero score on this item will be higher than for any other category. As long as a person thinks that the importance of this item for measuring hedonism is between [-0.502;0.414], the person will most probably select a response of "1," but need not do so. A person may choose "4," but such a response is much less likely than "1." Thus, unlike in CTT, the researcher can interpret the results in a probabilistic manner.

The items "emotional" and "full of action" possess a "regular" behavior in terms of increasing intersection parameters as the category increases; that is, $\delta_{i0} < \delta_{i1} < \delta_{i2} < \delta_{i3}$. This *monotonicity*

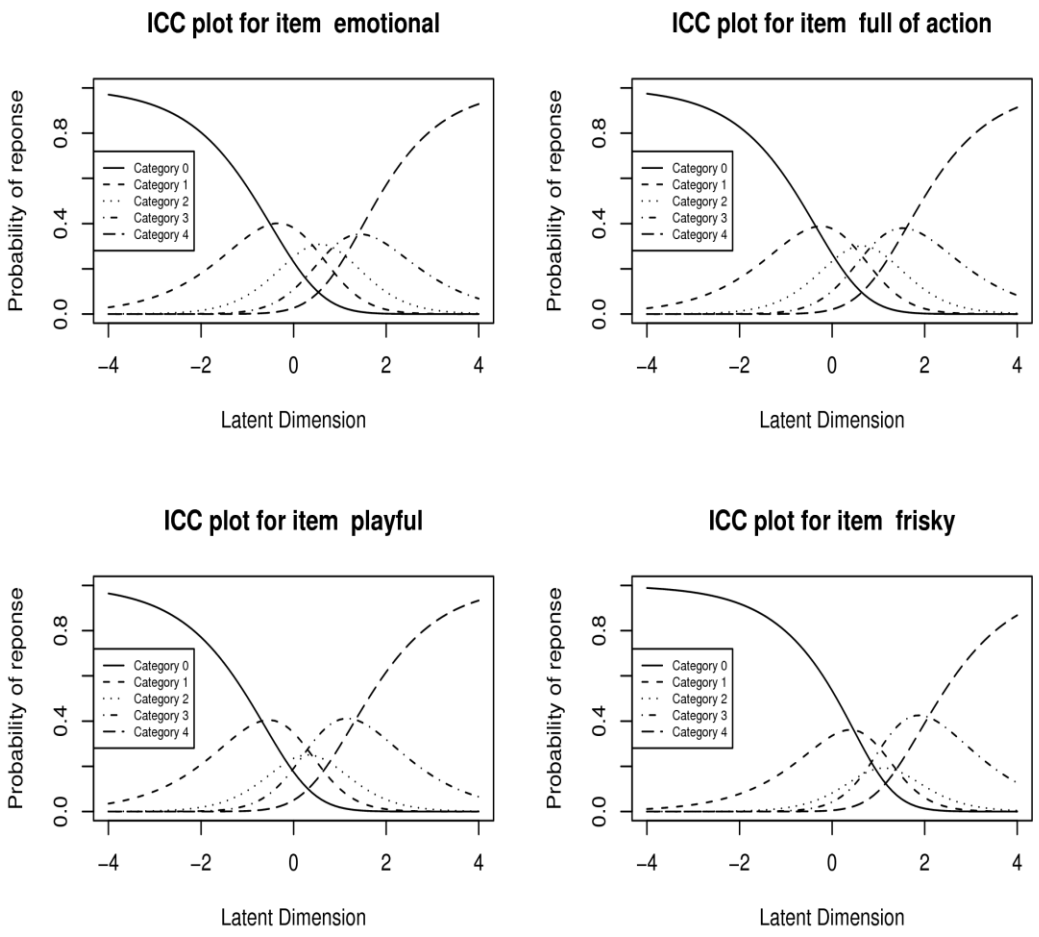

Figure 4. ICCs for the Five Categories

property is not given for all the other items (see bottom of Figure 5). It is especially striking that $\delta_{i2} < \delta_{i1}$ for the item "playful." This does not imply that there are not enough subjects with a score of 3, rather, it shows that conditional on the importance for hedonism score, the probability of a response in category 2 is lower throughout, compared to the responses in the other categories. This is a behavior that can be observed for neutral or middle categories frequently. The PCM assigns intervals on the one-dimensional latent trait to a score. These intervals must also be ordered; if this is not the case, it means that, in contrast to the assumption made when developing the items, category 3 cannot be mapped in this way. No interval is assigned mainly to this category, which suggests that something about this category is not in accordance with the ordinal ordering of the categories. For scale development, this indicates that the

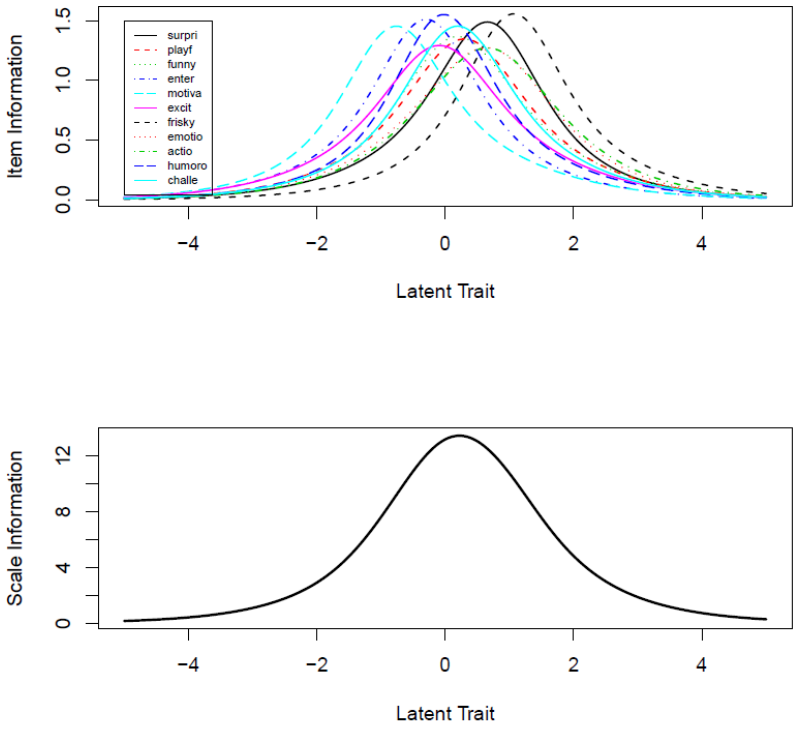

Figure 5. Information Curves for All Items (Top) and the Aggregated Information of the Scale (Bottom) as a Function of the Latent Trait Values

categories of items that display this behavior are not ordinal, but that there are four categories that are ordered, and that the middle category is different. One now has three choices: (1) either change the Likert scale to a 4-point scale with no middle category, since it measures something else, such as “undecided”; (2) use an IRT model that assumes the scores to be only nominal scaled, or (3) use a model that estimates the four “regular” categories as ordinal and the middle category as “nominal” [7].

To assess how well and where an item measures its latent trait values, IRT employs the concept of the *information of items*, which indicates how much information an individual item can give about certain latent trait values [76]. In doing so, we can see which region of the latent trait is measured well by which items, and which items are *redundant*. We can also add up these item information values into a joint information value if we want to compare different scales to measure hedonism. Figure 5 shows a plot of the item information (top) and scale information (bottom) as a function of the underlying latent trait

values. The attribute “exciting”, for example, measures the latent trait in an interval similar to the rest of the items, but has less information. Thus, if one wants to reduce the number of items further, this would be a potential candidate for removal. The attribute “motivating” is particularly important, because it has high information in the range on the latent trait, which stands for high importance in measuring hedonism. Conversely, the attribute “frisky” has the most information on the low values of the latent trait, which means it can be used to differentiate between low latent trait values.

4. DISCUSSION AND CONCLUSION

4.1 Theoretical Implications

[100], who use IRT to analyze the reliability of the leadership practices inventory, pointedly emphasized that “an instrument’s measurement precision is crucial for the quality of the inferences and decisions based on that instrument, whether the purpose is leader assessment in organizations or academic theory building” (p. 180). They further elaborate that wrong measurement invariably leads to wrong conclusions with far-reaching consequences. A further prominent example in this context is the still ongoing discussion about reflective vs. formative measurement , which was triggered in the social sciences in 2003 by Jarvis et al. [41] and then reached the IS community in 2007 by Petter et al. [60]. Both publications found that a substantial number of scales in the *existing* literature were actually misspecified. According to Peter et al., the misspecification level of publications in leading IS journal equals 30%. This illustrates that from time to time, paradigm shifts and critical “outside” evaluations of method are necessary to be able to increase the accuracy and validity of IS research results and conclusions. We suggest that IRT constitutes such an alternative that has the potential to uncover the shortcomings of CTT, which might otherwise go unnoticed as long as the existing measurement paradigm is not carefully scrutinized.

IS research frequently uses survey data to measure the interplay between technological systems and human beings and to create appropriate scales. However, most scales used in IS studies are based on a development process utilizing CTT, which suffers from major theoretical shortcomings. We advocate the

use of IRT as a viable alternative that overcomes the serious limitations of CTT models. Table 3 compares the applicability of CTT, FA, and IRT.

Table 3. Comparison of CTT and IRT Properties extended from Hambleton and Jones [35]

Point of Comparison	CTT	IRT
Model Type	Linear	Nonlinear
Scale Level of Item	Metric	Categorical
Level of Application	Item set	Individual item
Assumptions	Weak (easier to meet with data)	Strong (more difficult to meet with data)
Item-Ability Relationship	Not specified (usually linear function)	Item characteristic function
Ability Indicator (Range)	Test scores/estimated true score (restricted to range of raw scores)	Person parameter ($-\infty, +\infty$)
Invariance of Item & Person Statistics	No	Yes (if model holds)
Reliability/Internal Consistency	Estimated reliability	Model inherent
Assumptions Testable?	No	Yes

To address the advantages of IRT for IS research, we introduced the IRT paradigm of measurement in the IS context of hedonic websites and illustrated the practical applicability of a probabilistic framework in measuring latent constructs. We did so by means of polytomous Rasch models, in order to find attributes that are suitable for characterizing the hedonic aspects of websites. We derived and compared scales, with both the IRT and the CTT paradigms, and concluded that the scale derived with IRT not only had the same reliability and fewer items then the CTT scale, but also provided additional insights. Namely, IRT provides more information about the individual scale and its items, and embeds the scale construction process and the derived scholarly results into a strong theoretical and epistemological context of measurement. The IRT analysis not only allows for probabilistic statements about an individual's answering behavior, but also indicates (a) how well the expression of the latent construct subjects can be assessed, (b) how well the overall latent construct can be assessed, and (c) how the individual items scale the individuals.

Contrary to popular belief amongst many social researchers, IRT is fairly easy to perform with modern software packages. We used the open source software R to illustrate how to construct and test a scale that can be used to measure hedonic IS. This paper should help IS researchers to correctly interpret

the results. IRT is useful at virtually every stage of survey research. First, it can help with scale development and identify those items which do not carry much information. Second, it can help to better interpret the results and even to assess the suitability of different scale levels. As we have shown, it might even turn out a 4-point scale is preferable to a 5-point scale. This interpretation cannot be concluded with CTT. Third, IRT might help to overcome widespread misconceptions regarding the quality of scales. Cronbach's α, for example, is a frequently used indicator to measure the reliability of a scale and, most likely, one of the most misunderstood tests in social science research. Apart from the fact that it measures internal consistency rather than reliability, researchers frequently refer to Nunnally [58] who indicated .7 might be an acceptable coefficient. This does not mean, however, that better values necessarily indicate superior performance. In fact, previous literature recommends an upper value of .9 [32], with values above that level indicating redundant items. IRT can help to detect these items in the scale construction process.

By correctly applying this method, new insights about the content domain of frequently used constructs can be gained. Additionally, it is a powerful methodology for developing and testing new constructs. Another promising approach for future IRT applications in IS lies in the development and implementation of multi-dimensional IRT models, which map items and people simultaneously onto multiple correlated dimensions [96] or allow for measuring change over time [37]. We are quick to emphasize, however, that IRT is not always preferable. Although IRT is generally regarded to be superior to CTT for measurement purposes in behavioral science, the combination of both approaches is particularly powerful [8]. To date, there exists a dearth of studies that compare results gained from IRT and CTT. Such studies might help to shed light on what the differences actually are and how those might influence theory development.

Because IRT is a measurement paradigm, further research should account for the nomological context and the theoretical framework in which the respective constructs are being used; this is similar to suggestions from Burton-Jones and Straub [11] to operationalize constructs according to the specific

hypotheses or theory. We hope that our explanation and demonstration of the usefulness of IRT in an IS context further inspires such research.

4.2 Implications to Practice

Public and private organizations need reliable and effective tools to measure a wide variety of internal and external key indicators. Examples include work and financial performance, job and customer satisfaction, brand equity, and technology adoption. Furthermore, a wide variety of moderating and mediating variables, including extrinsic and intrinsic motivation, hedonic motives, usefulness and ease of use are frequently included in questionnaires. IRT not only bears the potential to make these surveys more efficient and less time-consuming, which is due to exclusion of redundant variables, but also allows for the application of new measurement paradigms that may offer considerable advantages as we have shown in previous sections. Often being seen as solely useful for psychological assessments (to date the majority of IRT studies are indeed published in Psychology journals), several studies from fields such as Marketing, Finance, Engineering and Business Administration in general are starting to demonstrate the usefulness of this approach. A famous example of IRT outside of academia includes its application for the Programme for International Student Assessment (PISA), which is the worldwide study of the Organization for Economic Co-operation and Development (OECD) to assess pupils' school performance. The first PISA study was performed in 2000 and it was then repeated every three years with 510,000 students from 65 nations and territories participating in 2012 [59]. IRT in this case allows for standardized, cross-national and accurate measurement and continues to be the method of choice for this study.

In this manuscript, we present the technical details of IRT in a manner that should be more accessible to a general audience than those seen in Psychology publications. The source code, which can be found in the online Appendix B, shows how to perform an IRT study, which can be achieved comparatively easy with modern software packages. SAS, STATA, SPSS and EQSIRT, which are widely used within the industry and academia, all allow for the application of IRT. We used R [61], which is open

source software increasingly getting attention in both academia and amongst practitioners. In R, all packages needed to perform IRT can be installed on the fly and without extra cost. The theoretical background in combination with a step-by-step tutorial should make it easy for practitioners to successfully apply this powerful method. In Appendix C, we provide a basic example that allows readers to test an easy-to-understand IRT application. A detailed interpretation of the results can be found online in [12]. Finally, it was our goal to raise awareness amongst managers and practitioners that alternative measurement approaches exist and to illustrate how to interpret the findings from studies and reports applying IRT.

4.3 Limitations of IRT

IRT certainly has disadvantages and limitations, which we briefly discuss here to conclude this manuscript. It is important that IS researchers also consider the downsides of IRT so that they do not blindly rethink the measurement of categorical indicators without understanding the risks.

First, IRT operates mainly on the item level. CTT offers the analysis of a scale on the set of item level, which IRT does not address. Consequently, if one wants to know the properties of the overall scale, CTT is a viable option. IRT can help in ensuring that CTT can actually be used for a set of items. For example, if a Rasch model holds, then the raw score of a set of items is sufficient. CTT concepts like validity and reliability can then be used for a set of items that has been constructed with IRT, and the test characteristic function (the sum of all item characteristics) connects the Θ from IRT to the true score of CTT. Importantly, IRT models are not ideal when the items are actually *metric*, such as with a true continuous response scale.

Second, IRT models are more complex than CTT models. They are more difficult to fit, and thus their parameters may only be estimated insufficiently, or a higher number of observations are needed to achieve sufficient accuracy of the estimates. Likewise, IRT models can be more difficult to learn and interpret for practitioners, which particularly applies to non-Rasch models.

Third, factor analysis methods can be combined with latent regression models to conduct

powerful analyses via SEM, which approach essentially correlates/regresses among the latent traits defined by the measurement models. Due to the considerably higher complexity of IRT models, all attempts at providing a similarly solid foundation for SEM with IRT measurement that we know of have fallen short, but there are promising developments [e.g., 56, 79].

Fourth, IRT models are by no means free of assumptions. For example, most assume some form of independence of the items, conditional on the latent trait, or assume a one-dimensional latent trait. If the assumptions behind the IRT model are not met, then the properties expected from the model and the inferences based on the model are not accurate—just as we previously criticized CTT for.

Fifth, choosing the right model is still a bit problematic. Due to the number of IRT models in existence, the sheer number of possibilities for modeling the data is huge.

4.4 Conclusion and Further Research

We have made the case that IS researchers have overwhelmingly favored CTT use for measurement development, even though there are downsides to this approach. CTT has a number of shortcomings when applied to categorical item scales, including the assumption of linearity, the difficulty of estimating the true score, and the sample dependence of the parameter estimates. To address these issues, we presented IRT as a collection of viable alternatives for measuring continuous latent variables by means of categorical indicators. IRT can overcome the serious limitations of CTT by offering: nonlinear relationships and appropriate estimation of the true score, possible sample independence of the parameters, and model-based procedures for selecting items that are in accordance with a desired model. IRT also generalizes concepts such as reliability or internal consistency, and allows a researcher to acquire a deep understanding of the measurement process. We conclude that a better (i.e., more precise) measurement increases the overall validity of the constructs being used in IS research and hence the explanatory power of both theory building and theory testing increases. We provided an empirical demonstration of creating a hedonic IS scale using the CTT approach and IRT approach. The results illustrate how IRT can be successfully applied in IS research with advantages over the traditional CTT

approach. Previous research has mostly favored one paradigm over the other. Hence, that further research is needed that actually compares the outcomes of CTT and IRT analyses and their implications for theory.

REFERENCES

1. Adams, R.J.; Wu, M.L.; and Carstensen, C.H. Application of multivariate Rasch models in international large-scale educational assessments. In M. Davier, and C.H. Carstensen (eds.), *Multivariate and Mixture Distribution Rasch Models: Extensions and Applications*. New York, NY: Springer, 2007, pp. 271-280.
2. Agarwal, R. and Karahanna, E. Time flies when you're having fun: Cognitive absorption and beliefs about information technology usage. *MIS Quarterly*, 24, 4 (2000), 665-694.
3. Agresti, A. *Analysis of Ordinal Categorical Data*, 2nd ed. Hoboken, NJ: John Wiley & Sons, Inc., 2010.
4. Alvarez, P.; Lopez-Rodriguez, F.; Canito, J.L.; Moral, F.J.; and Camacho, A. Development of a measure model for optimal planning of maintenance and improvement of roads. *Computers & Industrial Engineering*, 52, 3 (2007), 327-335.
5. Andrich, D. *Rasch Models for Measurement*. Newbury Park, CA: Sage, 1988.
6. Bartholomew, D.J. *Measuring Intelligence: Facts and Fallacies*. Cambridge, U.K.: Cambridge University Press, 2004.
7. Bartholomew, D.J. and Knott, M. *Latent Variable Models and Factor Analysis*. London, UK: Hodder Arnold, 1999.
8. Bechger, T.M.; Maris, G.; Verstralen, H.H.; and Béguin, A.A. Using classical test theory in combination with item response theory. *Applied Psychological Measurement*, 27, 5 (2003), 319-334.
9. Birnbaum, A. Some latent trait models. In F.M. Lord, and M.R. Novick (eds.), *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Wesley, 1968, pp. 395-479.
10. Borsboom, D. The attack of the psychometricians. *Psychometrika* 71, 3 (2006), 425-440.
11. Burton-Jones, A. and Straub, W. Reconceptualizing system usage: An approach and empirical test. *Information Systems Research*, 17, 3 (2006), 228-246.
12. Cadwell, J. Item response theory: Developing your intuition. (2012), Date last accessed: October 10, 2015, retrieved from http://joelcadwell.blogspot.co.at/2012/09/item-response-theory-developing-your.html
13. Chin, W.W.; Junglas, I.; and Roldán, J.L. Some considerations for articles introducing new and/or novel quantitative methods to IS researchers. *European Journal of Information Systems*, 21, 1 (2012), 1-5.
14. Cronbach, L.J. Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 1951 (1951), 297-334.
15. Dawes, J.G. Do data characteristics change according to the number of scale points used ? An experiment using 5 point, 7 point and 10 point scales. *International Journal of Market Research*, 51, 1 (2008), 61-104.
16. de Boeck, P. and Wilson, M. *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach*. New York, NY: Springer, 2004.
17. Dekleva, S. and Drehmer, D. Measuring software engineering evolution: A Rasch calibration. *Information Systems Research*, 8, 1 (1997), 95–104.
18. Deng, L.; Turner, D.E.; Gehling, R.; and Prince, B. User experience, satisfaction, and continual usage intention of IT. *European Journal of Information Systems*, 19, 1 (2010), 60-75.
19. Dittrich, R.; Francis, B.; Hatzinger, R.; and Katzenbeisser, W. A paired comparison approach for the analysis of sets of Likert-scale responses. *Statistical Modelling*, 7, 1 (2007), 3-28.
20. Edelen, M.O. and Reeve, B.B. Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research*, 16, 1 (2007), 5-18.

21. Embretson, S.E. and Reise, S. *Item Response Theory for Psychologists*. Mahwah, NJ: Lawrence Erlbaum, 2000.
22. Ewing, M.; Salzberger, T.; and Sinkovics, R.R. An alternate approach to assessing cross-cultural measurement equivalence. *Journal of Advertising*, 34, 1 (2005), 17-36.
23. Fischer, G.H. *Einführung in die Theorie psychologischer Tests [Introduction to Mental Test Theory]*. Bern, Germany: Huber, 1974.
24. Fischer, G.H. and Formann, A.K. Some applications of logistic latent trait models with linear constraints on the parameters. *Applied Psychological Measurement*, 6, 1982 (1982), 397-416.
25. Fox, J. polycor: Polychoric and Polyserial Correlations R package version 0.7-7. (2009), Date last accessed: November 6, 2012, retrieved from http://cran.r-project.org/web/packages/polycor/
26. Fox, J. sem: Structural Equation Models. R package version 0.9-16. (2009), Date last accessed: April 18, 2012, retrieved from http://CRAN.R-project.org/package=sem
27. Ganglmair-Wooliscroft, A. A comparison of affective response to consumption in two contexts'. *der markt: International Journal of Marketing*, 46, 1-2 (2007), 36-49.
28. Ganglmair, A. and Lawson, R. Advantages of Rasch modelling for the development of a scale to measure affective response to consumption. *European Advances in Consumer Research*, 6, 2003 (2003), 162-168.
29. Ganglmair, A. and Lawson, R. Measuring affective response to consumption using Rasch modelling. *Journal of Customer Satisfaction, Dissatisfaction and Complaining Behavior*, 16, 2003 (2003), 198-210.
30. Glas, C. and Verhelst, N. Tests of fit for polytomous Rasch models. In G.H. Fischer, and I.W. Molenaar (eds.), *Rasch Models. Their Foundation, Recent Developments and Applications*. New York, NY: Springer, 1995, pp. 325-352.
31. Göb, R.; McCollin, C.; and Ramalhoto, M. Ordinal methodology in the analysis of Likert scales. *Quality & Quantity*, 41, 5 (2007), 601-626.
32. Green, S.B.; Lissitz, R.W.; and Mulaik, S.A. Limitations of coefficient alpha as an index of test dimensionality. *Educational and Psychological Measurement*, 37, 4 (1977), 827–838.
33. Gulliksen, H. *Theory of Mental Tests*. Hoboken, NJ: Wiley, 1950.
34. Hambleton, R.; Swaminathan, H.; and Rogers, H. *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage Publications, Inc., 1991.
35. Hambleton, R.K. and Jones, R.W. Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement*, 12, 3 (1993), 38-47.
36. Hart, M.C. Improving the discrimination of SERVQUAL by using magnitude scaling. In G.K. Kanji (ed.), *Total Quality Management in Action*. London: Chapman & Hall, 1996, pp. 267-270.
37. Hatzinger, R. and Rusch, T. IRT models with relaxed assumptions in eRm: A manual-like instruction. *Psychology Science Quarterly*, 51, 1 (2009), 87-120.
38. Hooper, D.; Coughlan, J.; and Mullen, M.R. Structural equation modelling: Guidelines for determining model fit. *Electronic Journal of Business Research Methods*, 6, 1 (2008), 53–59.
39. Huang, E. The acceptance of women-centric websites. *Journal of Computer Information Systems*, 45, 4 (2005), 75-83.
40. Jamieson, S. Likert scales: How to (ab)use them. *Medical Education*, 38, 12 (2004), 1217-1218.
41. Jarvis, C.B.; Mackenzie, S.B.; and Podsakoff, P.M. A critical review of construct indicators and measurement model misspecification in marketing and consumer research. *Journal of Consumer Research*, 30, 2 (2003), 199-218.
42. Likert, R. A technique for the measurement of attitudes. *Archives of Psychology*, 22, 140 (1932), 1-55.
43. Lin, C.-P. and Bhattacherjee, A. Extending technology usage models to interactive hedonic technologies: A theoretical model and empirical test. *Information Systems Journal*, 20, 2 (2010), 163-181.
44. Lord, F.M. *Applications of Item Response Theory to Practical Testing Problems*. Mahwah, NJ: Lawrence Erlbaum, 1980.

45. Lord, F.M. and Novick, M.R. *Statistical Theories of Mental Test Scores*. Reading, MA: Addison Wesley, 1968.
46. Lowry, P.B.; Gaskin, J.; and Moody, G.D. Proposing the multimotive information systems continuance model (MISC) to better explain end-user system evaluations and continuance intentions. *Journal of the Association for Information System*, forthcoming, (2015),
47. Lowry, P.B.; Gaskin, J.; Twyman, N.W.; Hammer, B.; and Roberts, T.L. Taking 'fun and games' seriously: Proposing the hedonic-motivation system adoption model (HMSAM). *Journal of the Association for Information Systems*, 14, 11 (2013), 617-671.
48. Lubke, G. and Muthen, B. Applying multigroup confirmatory factor models for continuous outcomes to Likert scale data complicates meaningful group comparisons. *Structural Equation Modeling: A Multidisciplinary Journal*, 11, 4 (2004), 514-534.
49. Luce, R.D.; Krantz, D.H.; Suppes, P.; and Tversky, A. *Foundations of Measurement*. III. New York, NY: Academic Press, 1990.
50. MacKenzie, S.B.; Podsakoff, P.M.; and Podsakoff, N.P. Construct measurement and validation procedures in MIS and behavioral research: Integrating new and existing techniques. *MIS Quarterly*, 35, 2 (2011), 293-334.
51. Mair, P. and Hatzinger, R. CML based estimation of extended Rasch models with the eRm package in R. *Psychology Science*, 49, 2007 (2007), 26-43.
52. Mair, P. and Hatzinger, R. Extended Rasch modeling: The eRm package for the application of IRT models in R. *Journal of Statistical Software*, 20, 9 (2007), 1-20.
53. Masters, G.N. A Rasch model for partial credit scoring. *Psychometrika*, 47, 1982 (1982), 149-174.
54. Melas, C.D.; Zampetakis, L.A.; Dimopoulou, A.; and Moustakis, V.S. The significance of attitudes towards evidence-based practice in information technology use in the health sector: An empirical investigation. *Behaviour & Information Technology*, 33, 12 (2014), 1248–1260.
55. Moore, G.C. and Benbasat, I. Development of an instrument to measure the perceptions of adopting an information technology innovation. *Information Systems Research*, 2, 3 (1991), 192-222.
56. Muthen, B. A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, 49, 1 (1984), 115-132.
57. Nicolosi, M.; Grassi, S.; and Stanghellini, E. Item response models to measure corporate social responsibility. *Applied Financial Economics*, 24, 22 (2014), 1449–1464.
58. Nunnally, J.C. *Psychometric Theory*, 2nd ed. New York, NY: McGraw-Hill, 1978.
59. OECD. *Pisa 2012 Technical Report*. Paris, France: Organisation of Economic Co-Operation and Development, 2014.
60. Petter, S.; Straub, D.; and Rai, A. Specifying formative constructs in information systems research. *MIS Quarterly*, 31, 4 (2007), 623-656.
61. R Core Team. R: A language and environment for statistical computing. (2014), Date last accessed: September 23, 2015, retrieved from http://www.R-project.org
62. R Development Core Team. R: A Language and Environment for Statistical Computing. (2007), Date last accessed: April 18, 2012, retrieved from http://www.r-project.org/
63. Ramsay, J.O. Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika*, 56, 4 (1991), 611-630.
64. Rasch, G. *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen, Denmark: Danish Institute for Educational Research, 1960.
65. Rasch, G. On general laws and the meaning of measurement in psychology. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, IV. Berkeley, CA: University of California Press, 1961, pp. 321-333.
66. Rasch Org. Definition of objective measurement. (2000), Date last accessed: October 22, 2012, retrieved from http://www.rasch.org/define.htm
67. Raykov, T. and Calantone, R. The utility of item response modeling in marketing research. *Journal of the Academy of Marketing Science*, 42, 4 (2014), 337–360.

68. Reise, S.P. and Revicki, D. *Handbook of Item Response Theory Modeling: Applications to Typical Performance Assessment*. New York, NY: Taylor & Francis, 2014.
69. Revelle, W. psych: Procedures for psychological, psychometric, and personality research. R package version 1.0-67. (2009), Date last accessed: April 18, 2012, retrieved from http://CRAN.R-project.org/package=psych
70. Revelle, W. and Rocklin, T. Very simple structure: An alternative procedure for estimating the optimal number of interpretable factors. *Multivariate Behavioral Research*, 14, 1979 (1979), 403-414.
71. Salzberger, T. Scientific Measurement of Latent Variables in Marketing Research: An Alternative Framework. Postdoctoral Lecture Qualification. Vienna, Switzerland: Vienna University of Economics and Business Administration, 2007.
72. Salzberger, T.; Holzmüller, H.; and Souchon, A. Advancing the understanding of construct validity and cross-national comparability: Illustrated by a five-country study of corporate export information usage. In R.R. Sinkovics, and P.N. Ghauri (eds.), *New Challenges to International Marketing (Advances in International Marketing)*, 20. Bingley, UK: Emerald Group Publishing Limited, 2009, pp. 321-360.
73. Salzberger, T. and Koller, M. Towards a new paradigm of measurement in marketing. *Journal of Business Research*, in press, (2012),
74. Salzberger, T.; Newton, F.J.; and Ewing, M.T. Detecting gender item bias and differential manifest response behavior: A Rasch-based solution. *Journal of Business Research*, 67, 4 (2014), 598–607.
75. Salzberger, T. and Sinkovics, R.R. Reconsidering the problem of data equivalence in international marketing research. *International Marketing Review*, 23, 4 (2004), 390-417.
76. Samejima, F. Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, 35, 1 (1970), 139-139.
77. Shang, R.-A.; Chen, Y.-C.; and Shen, L. Extrinsic versus intrinsic motivations for consumers to shop on-line. *Information and Management*, 42, 3 (2005), 401-413.
78. Sijtsma, K. On the use, the misuse, and the very limited usefulness of Cronbach's Alpha. *Psychometrika*, 74, 1 (2009), 107-120.
79. Skrondal, A. and Rabe-Hesketh, S. *Interdisciplinary Statistics: Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*. Boca Raton, FL: Chapman and Hall/CRC, 2004.
80. Smith, R.M. Fit analysis in latent trait measurement models. In E.S. Smith, and R.M. Smith (eds.), *Introduction to Rasch Measurement*. Maple Grove, MN: JAM Press, 2004, pp. 73-92.
81. Spearman, C. General intelligence objectively determined and measured. *American Journal of Psychology*, 15, 202 (1904),
82. Spearman, C. *The Abilities of Man*. London, U.K.: Macmillan, 1927.
83. Stevens, S.S. On the theory of scales of measurement. *Science*, 103, 1946 (1946), 667-680.
84. Stevens, S.S. Mathematics, measurement and psychophysics. In S.S. Stevens (ed.), *Handbook of Experimental Psychology*. New York, NY: John Wiley & Sons, 1951, pp. 1-49.
85. Thurstone, L.L. A method of scaling psychological and educational tests. *Journal of Educational Psychology*, 16, 1925 (1925), 433-451.
86. Thurstone, L.L. *The Measurement of Values*. Chicago, IL: The University of Chicago Press, 1959.
87. Treiblmaier, H. and Filzmoser, P. Exploratory factor analysis revisited: How robust methods support the detection of hidden multivariate data structures in IS research. *Information & Management*, 47, 4 (2010), 197–207.
88. Treiblmaier, H. and Filzmoser, P. Benefits from using continuous rating scales in online survey research. Presented at *The International Conference on Information Systems (ICIS)*, Shanghai, China, 2011.
89. Tuerlinckx, F. and Wang, W. Models for polytomous data. In P. de Boeck, and M. Wilson (eds.), *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach*. New York, NY: Springer, 2004, pp. 75-110.

90. Tukey, J.W. Data analysis and behavioral science or learning to bear the quantitative burden by shunning badmandments. In L.V. Jones (ed.), *The Collected Works of John W. Tukey*, III. Belmont: Wadsworth, 1961, pp. 391–484.
91. van Alphen, A.; Halfens, R.; Hasman, A.; and Imbos, T. Likert or Rasch? Nothing is more applicable than good theory. *Journal of Advanced Nursing*, 20, 1 (1994), 196-201.
92. van der Heijden, H. User acceptance of hedonic information systems. *MIS Quarterly*, 28, 4 (2004), 695-704.
93. van der Linden, W.J. and Hambleton, R.K. *Handbook of Modern Item Response Theory*. New York, NY: Springer, 1996.
94. Venkatesh, V. Determinants of perceived ease of use: Integrating control, intrinsic motivation, and emotion into the technology acceptance model. *Information Systems Research*, 11, 4 (2000), 342-365.
95. Vincent, D.F. The origin and development of factor analysis. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 2, 2 (1953), 107-117.
96. von Davier, M. and Carstensen, C.H. *Multivariate and Mixture Distribution Rasch Models: Extensions and Applications*. New York, NY: Springer, 2007.
97. Wakefield, R.L. and Whitten, D. Mobile computing: A user study on hedonic/utilitarian mobile device usage. *European Journal of Information Systems*, 15, 3 (2006), 292-300.
98. Wang, L. and Finn, A. A consumer-based brand equity study for small market share brands. *Market & Social Research*, 22, 2 (2014), 6–14.
99. Weiss, D.J. and Davison, M.L. Test theory and methods. *Annual Review of Psychology*, 32, 1981 (1981), 629-658.
100. Zagorsek, H.; Stough, S.J.; and Jaklic, M. Analysis of the reliability of the leadership practices inventory in the item response theory framework. *International Journal of Selection and Assessment*, 14, 2 (2006), 180-191.
101. Zhang, P. and Li, N. Love at first sight or sustained effect? The role of perceived affective quality on users' cognitive reactions to information technology. Presented at *25th International Conference on Information Systems*, Washington, DC, 2004, pp. 283-295.

APPENDIX A. RESULTS OF THE CONFIRMATORY FACTOR ANALYSIS

	One factor solutions				Two factor solution			
	Factor 1 before selection		Factor 1 after selection		Factor1		Factor2	
Item	**Loading**	**SE**	**Loading**	**SE**	**Loading**	**SE**	**Loading**	**SE**
surprising	0.696	0.052	0.715	0.052	0.715	0.052	-	-
intriguing	0.563	0.055	0.557	0.056	0.553	0.056	-	-
inspiring	0.589	0.055	0.563	0.055	-	-	0.656	0.056
playful	0.694	0.052	0.702	0.052	0.703	0.052	-	-
animated	0.659	0.053	0.650	0.054	0.651	0.054	-	-
multimedia based	0.506	0.056	-	-	-	-	0.585	0.057
funny	0.687	0.053	0.697	0.052	0.707	0.052	-	-
entertaining	0.762	0.051	0.767	0.051	0.770	0.050	-	-
provocative	0.537	0.056	0.564	0.055	0.561	0.056	-	-
motivating	0.591	0.055	0.561	0.056	-	-	0.696	0.055
beautiful	0.504	0.056	-	-	-	-	0.573	0.058
exciting	0.723	0.052	0.717	0.052	0.715	0.052	-	-
coltish	0.767	0.050	0.785	0.050	0.795	0.050	-	-
modern	0.423	0.058	-	-	-	-	0.496	0.059
emotional	0.703	0.052	0.697	0.052	0.695	0.052	-	-
colorful	0.619	0.054	0.618	0.054	0.622	0.054	-	-
full of action	0.716	0.052	0.735	0.051	0.734	0.052	-	-
humorous	0.742	0.051	0.748	0.051	0.748	0.051	-	-
challenging	0.667	0.053	0.670	0.053	0.663	0.053	-	-
interactive	0.335	0.059	-	-	0.316	0.059	-	-
customized	0.399	0.058	-	-	-	-	0.433	0.061
personalized	0.424	0.058	-	-	-	-	0.513	0.059
tasteful	0.370	0.059	-	-	-	-	0.550	0.059
suitable for children	0.478	0.057	0.475	0.057	0.474	0.057	-	-
creative	0.580	0.055	0.544	0.056	-	-	0.698	0.055
InterfactCorr					0.765			
Chi-Square (df)	1264.6 (275)		578.71 (135)		1106.8 (274)			
CFI	0.725		0.835		0.768			
RMSEA	0.111		0.106		0.102			
NFI	0.676		0.797		0.717			
NNFI	0.700		0.814		0.747			
SRMR	0.008		0.06		0.071			
Cronbach's alpha	0.91		0.91		0.78		0.90	

ONLINE APPENDIX B. SOURCE CODE

```
install.packages(c('eRm','psych','polycor','sem','foreign')) #install add on packages to base R;
we used version 0.15-1 of eRm, 1.2.1 of psych and  0.8-49 of foreign. Should updates to the
packages break the code below the packages can be installed by

#packageurls <- c('http://cran.r-project.org/src/contrib/Archive/eRm/eRm_0.15-
1.tar.gz','http://cran.r-
project.org/src/contrib/Archive/psych/psych_1.2.1.tar.gz','http://cran.r-
project.org/src/contrib/Archive/psych/foreign_0.8-49.tar.gz')
#install.packages(packageurls, repos=NULL, type='source')

#load packages
library('eRm')
library('psych')
library('foreign')
source('entertainment/ipmap2.R')      #this is script file that we use to generate a nicer plot
of the IP Map

#####################IRT ANALYSIS
#
#open irt.RData
#datana... raw data with NA
#data ... matrix with raw data
#X ... reduced entertainment matrix

#-------read data----------
datana <- read.spss('entertainment/Entertainment_ordinal.sav', use.value.labels = FALSE,
to.data.frame = TRUE)
xna <- datana[,c(8,9,13,15,16,19,21,23,25,27,28,29,30,33,34,35,42,44,46,47,14,18,20,31,41,43)]
xna[xna == -1] <- NA
tfvec <- apply(xna, 1, function(x) any(is.na(x)))
X <- xna[!tfvec,]                                    #NA's eliminated
#------end read data------


#---------------------------- Fit Assessment and Scale Development ---------------------

XS <- X
res <- PCM(X) #fit the PCM
summary(res)  #summary of the fit

pres <- person.parameter(res) #get the person parameters
lrres <- LRtest(res) #get the LR test of item fit
lrres   #fit is bad

#Why is the fit bad?
ifres <- itemfit(pres) #inspect items
ifres


plotGOF(lrres) #plot the problem in fit; we see that some items deviate from the 45degree line

#We have a bunch of items but the PCM does not fit them well as a whole; we therefore proceed by
a successive item elimination strategy of deleting a _single_ candidate item, refitting the
model, assessing fit, removing another item, refitting the model, assessing fit again, and so on
until the fit is acceptable.

X <- X[,-25]         #eliminate SCHLICHT
restmp <- PCM(X) #fit the PCM
prestmp <- person.parameter(restmp) #get the person parameters
lrrestmp <- LRtest(restmp) #get the LR test of item fit
lrrestmp   #fit is still bad

# We would now do this for each of the X listed subsequently. For brevity we do not list the
intermediate steps of
## restmp <- PCM(X)
## prestmp <- person.parameter(restmp)
## lrrestmp <- LRtest(restmp)
## lrrestmp
```

```
# after each item elimination. But since the item elimination is conditional, we do not delete
all the columns in one go in the procedural reproduciblity script (of course after the successive
elimination we know the items that should be deleted and could remove all of them in one go by
writing
## dropcols <- c("SCHLICHT","KINDERGE",...)
## X <- XS[,-dropcols]


X <- X[,-25]         #eliminate KINDERGE
X <- X[,-21]         #eliminate INTERAKT
X <- X[,-21]         #eliminate INDIVIDU
X <- X[,-21]         #eliminate PERSONAL
X <- X[,-7]          #eliminate MULTIMED
X <- X[,-14]         #eliminate MODERN
X <- X[,-19]         #eliminate GESCHMAC

#lrres <- LRtest(res)
#p-value:  0.002

X <- X[,-11]         #eliminate SCHOEN
X <- X[,-1]          #eliminate KREATIV
X <- X[,-8]          #eliminate PROVOZIE


#lrres <- LRtest(res)
#p-value:  0.179

X <- X[,-3]          #eliminate INSPIRIE
X <- X[,-2]          #eliminate FESSELND
X <- X[,-10]         #eliminate BUNT
X <- X[,-3]          #eliminate ANIMIERT


#The final results

resfinal <- PCM(X)
lrfinal <- LRtest(resfinal)
lrfinal
plotGOF(lrfinal) #plot the GOF test
trfinal <- thresholds(resfinal) #getting the thresholds
trfinal
presfinal <- person.parameter(resfinal)
fitfinal <- itemfit(presfinal)


#Simulating 20 random splits (of 2 and 3) of the data to obtain the posterior of the test
statistics for the final result
testsim <- matrix(ncol = 2, nrow = 20)
for(i in 1:20)
  {
  g2 <- sample(c(0,1), 291, replace = TRUE)
  g3 <- sample(c(0,1,2), 291, replace = TRUE)
  tes2 <- LRtest(resfinal, splitcr = g2)
  tes3 <- LRtest(resfinal, splitcr = g3)
  testsim[i,1] <- tes2$pvalue
  testsim[i,2] <- tes3$pvalue
  cat(i,'\n')
}

testsim #the result of the simulations


###Plotting
#we give corresponding English labels to the German labels
colstring <-
c('surprising','playful','funny','entertaining','motivating','exciting','frisky','emotional','ful
l of action','humorous','challenging')
colnames(resfinal$X) <- colstring

#Plotting the (customized) Item Person Map
```

```
rawscore <-rowSums(resfinal$X)  #raw score
tr <- as.matrix(trfinal$threshtable[[1]]) #category thresholds in matrix
rownames(tr) <- colstring
theta <- presfinal$pred.list[[1]]$y[rawscore+1] #estimated theta values

#Plotting the Item Person Map
ipmap2(tr, theta)


#Plotting the ICCs interactively
plotICC(resfinal, item.subset = c(8,9,2,7), col = 1, lty = 1:5, ylab = 'Probability of response',
cex = 0.5)

#plotting  ICC of item 9
plotICC(resfinal, item.subset = 9, col = 1, lty = 1:5, ylab = 'Probability of response', ask =
FALSE, cex = 0.5)

#plotting  ICC of item 10
plotICC(resfinal, item.subset = 10, col = 1, lty = 1:5, ylab = 'Probability of response', ask =
FALSE)

##plotting ICC of item 2
plotICC(resfinal, item.subset = 2, col = 1, lty = 1:5, ylab = 'Probability of response', ask =
FALSE)

##plotting ICC of item 8
plotICC(resfinal, item.subset = 8, col = 1, lty = 1:5, ylab = 'Probability of response', ask =
FALSE)
```

ONLINE APPENDIX C. DEMONSTRATION SOURCE CODE

Note: this code is taken from Joel Cadwell (2012) and modified slightly
(http://joelcadwell.blogspot.co.at/2012/09/item-response-theory-developing-your.html)

In order to run in R the GenOrd and the ltm packages have to be installed:

```
install.packages('GenOrd')
install.packages('ltm')


#use GenOrd package to generate random data
library(GenOrd)
library('psych')


#probabilities for each brand test (location)
prob <- list(
  c(0.25),
  c(0.35),
  c(0.45),
  c(0.55),
  c(0.65)
  )


#slope for each logistic curve
loadings<-matrix(c(
  .6,
  .6,
  .6,
  .6,
  .6),
5, 1, byrow=TRUE)

#creates correlation matrix as input
cor_matrix<-loadings %*% t(loadings)
diag(cor_matrix)<-1

#generates 200 random ordinal observations
ord<-ordsample(n = 200, marginal = prob, Sigma = cor_matrix)

#calculates first principal component
library(psych)
principal(ord,nfactors=1)$value

library(ltm)
ord<-ord-1
descript(ord)

#likelihood ratio test
anova(rasch(ord), ltm(ord ~ z1))

#two-parameter logistic model
fit<-ltm(ord ~ z1)
summary(fit)

#item characteristic curves
plot(fit)

#calculates latent trait scores
pattern<-factor.scores(fit)
#constrains slopes to be equal
fit2<-rasch(ord)
plot(fit2)
summary(fit2)
scores2<-factor.scores(fit2)
```

REFERENCES

OECD, 2014. Pisa 2012 Technical Report. Organisation of Economic Co-Operation and Development, Paris, France.