NORGES HANDELSHØYSKOLE

Norwegian School of Economics

Bergen, Fall 2013

Measuring Marketing Constructs:
A Comparison of Three Measurement Theories

Master Thesis within the main profile of Marketing and Brand Management

Author: Shan Lin

Thesis Supervisor: Professor Einar Breivik

Abstract

A large number of new constructs are introduced into marketing. These new constructs are important in the development of marketing theories. The validation of these constructs are primarily based on a factor analytical framework (e.g., Churchill's paradigm), with a validation rationale found in Classical Test Theory. However as the limitations of Classical Test Theory are widely realized by many researchers, alternative measurement theories might provide better and more coherent guidelines for marketing researchers with regard to how to measure and validate marketing constructs. In this thesis I look at two alternative measurement theories, Item Response Theory and Generalizability Theory, in addition to Classical Test Theory. Both Item Response Theory and Generalizability Theory have recently become more important for marketing measurement. This thesis addresses how the constructs are measured based on these three theories and how these measurement theories differ from each other. After a detailed review of each theory, the theories are contrasted, especially in terms of construct validation. It is found that Classical Test Theory, Item Response Theory, and Generalizability Theory vary in how they address different measurement issues. They differ in terms of how constructs are defined, measured, and validated. However, the validation process employed for these three theories can only provide empirical evidence, or indications, for the construct validity but cannot provide evidence as to whether constructs exist or not. This will be a challenging research question for future research.

Preface

The topic of this thesis was initiated in January 2013, together with my supervisor, Professor Einar Breivik. As a result of long lasting interest towards marketing research, this thesis not only completes my master of science in marketing and brand management at Norwegian School of Economics, but also serves as a starting point towards my future PhD dissertation at the same institution. As more and more constructs are introduced into marketing, I attempt to address whether these newly proposed constructs match with the reality. If the concept of interest does not exist, the whole process of measurement would be meaningless. However, addressing the problem of existence is a daunting task that exceeds the scope of a master thesis. Thus, this thesis only addresses how three measurement theories validate a construct, and serves as an initial part of a more comprehensive research program. I will further extend the research regarding whether the concept of interest is real in my doctoral study.

Working with the thesis has been inspiring, although challenging, both in terms of what I learned about the topic and what I have to learn in the future. In addition, I acknowledge that the theoretical frameworks that I now possess, together with my enhanced understanding of the topic, will be of great relevance to my future research work.

Furthermore, I would like to thank my supervisor, Einar Breivik, whose advice and support from the initial to the final level enabled me to develop an understanding of the subject. Lastly, I offer my regards and blessings to all of those who supported me in any respect during the completion of the thesis.

Shan Lin

Bergen, 30th July 2013

Table of Contents

1. Introduction

There is no shortage of constructs in the field of marketing. As most vital research fields, marketing is also characterized by a growth in the number of constructs. To illustrate a process suggesting such a development, we can look at brand loyalty. Since loyalty was first defined there has been a great deal of debate about the construct. Early marketing research (Ehrenberg, 1972; 1988) conceptualized loyalty as a behavioral outcome equal to repurchase. Hence, by observing repurchase patterns one also examined brand loyalty. However, later researchers (see e.g., Day, 1969; Jacoby and Kyner, 1973; Jacoby and Chestnut, 1978) questioned this conceptualization of loyalty. Their main rationale for questioning the use of repurchase alone as a measure of brand loyalty was that without understanding the factors underlying repeat purchase how could we tell that this repurchasing behavior really was an indicator of loyalty. What if consumers were repurchasing the brand because it was the least expensive or the one with highest objective quality? What would then happen if another producer offered a less expensive variant or one with higher quality? Hence, they also introduced the attitudinal component, commitment, as a part of brand loyalty (Jacoby and Chestnut, 1978).

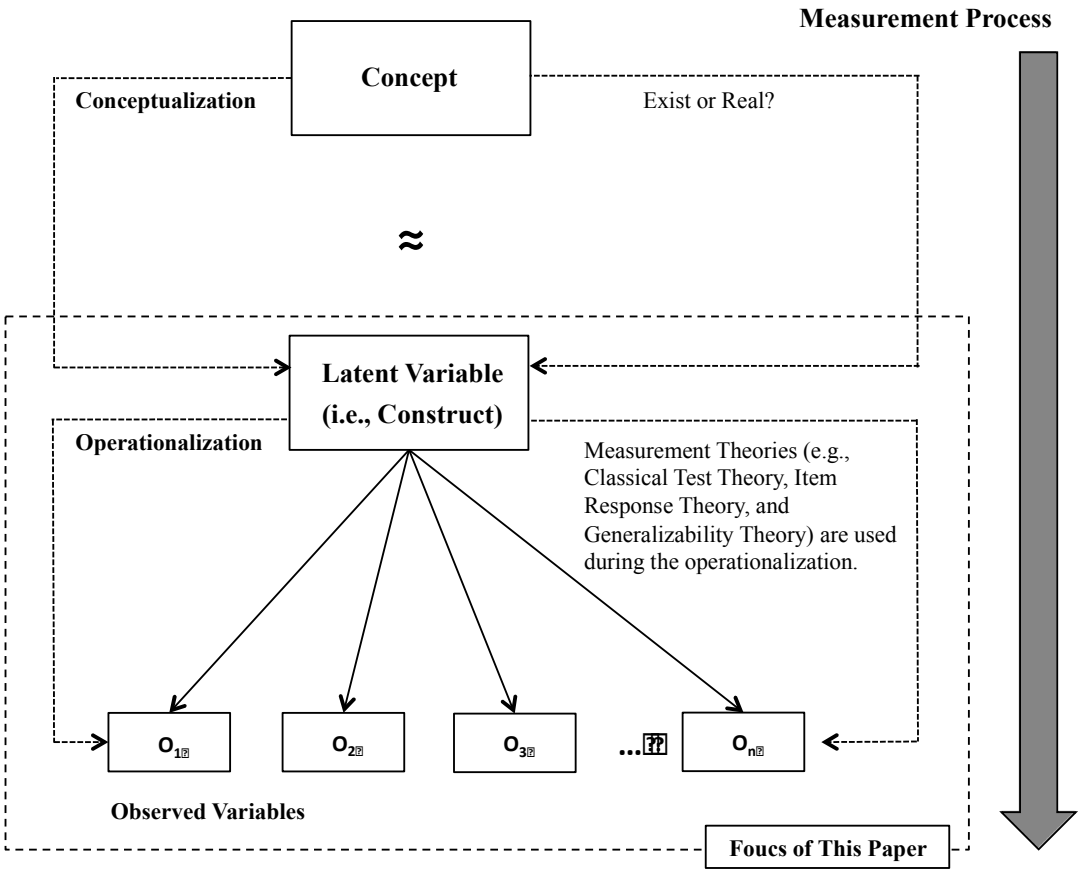
Afterwards, other researchers (Meyer and Allen, 1991) have split the commitment into affective, continuance (i.e., calculative), and normative components reflecting consumer's desire, need, and obligation respectively. Some researchers expanded the construct, loyalty, further to include a cognitive component reflecting consumers' brand beliefs (Bloemer et al., 1998; 1999), and other types of behavior apart from repurchase. This process represents an expansion from just looking at repurchase when one was addressing brand loyalty to a situation where one look at several different constructs when addressing brand loyalty. The common denominator of these new constructs is that they all are latent, they cannot be directly observed[1]. Similar trends can be found in all areas of marketing, such as the conceptualization of brand equity and relationship marketing.

[1] Even repurchase contains some latent parts (e.g., unreliability, missed measurements, etc.), but certainly the measurement of repurchase should be less ambiguous as compared to commitment or brand beliefs.

Latent constructs represent a challenge for marketing in terms of measurement. Perhaps the most basic question would be: do these latent constructs really exist? Do they represent some sort of reality? How could marketers ensure measurement validity? How could they be sure that the constructed latent variable is a valid representation of the concept of interest? How could they be confident in the evaluation procedures and measurement procedures they use to validate the concept?

The measurement process should address two basic questions. The first question would address whether we can capture a concept of interest by a latent variable (or latent variables) and the second question would address whether the procedures we utilize can link relevant observations to measure this latent variable (see Figure 1). If the latent variable is not properly conceptualized, it would provide limited explanatory value. Furthermore, a prerequisite for proper conceptualization would be that the concept of interest actually does exist. If not, the idea of measurement would be meaningless. Thus, construct validation should begin with verifying the existence of the concept of interest.

Figure 1: The framework of construct validation



Although the verification of construct existence is of vital importance, there are few guidelines and criteria available to establish existence. Hence a comprehensive treatment of this issue exceeds the scope of this thesis. In fact, I hope I will look more into this issue in future work with my dissertation. Instead, I will focus on the measurement theories in this thesis and thus for the most part assume that the constructs exist.

Given the existence of a concept or construct, marketing researchers need a measurement theory to capture the relation between the construct and its observables. Through interpreting the corresponding test scores, they would empirically assess whether the observed variables measure the construct reliably and validly. The dominant paradigm used for measuring latent constructs in marketing has been the factor analytic approach. The underlying rationale for the factor analytic approach can be found in Classical Test Theory (Lord and Novick, 1968). However, there are also other measurement approaches that can be found in marketing, such as Item Response Theory and Generalizability Theory. These approaches represent different perspectives on how to assess measurement of existing constructs. In this paper I will provide a detailed review of these three measurement theories and outline the different perspectives on how constructs are measured and validated based on the three different perspectives.

Research Problem

Recently, a rapid growth of new constructs can be found in the marketing literature. For instance the conceptualizations of brand loyalty (e.g., Knox and Walker, 2001; Punniyamoorthy and Raj, 2007; McMullan and Gilmore, 2003, 2008; Velázquez et al., 2011; etc.), brand personality (Grohmann, 2009; Geuens et al., 2009; etc.), brand experience (e.g., Schmitt et al., 2009; Kim et al., 2012), brand relationship quality (Fournier, 1998; Aggarwal, 2004; Chang and Chieng, 2006) and so forth all represent examples of recently proposed constructs. Furthermore, researchers continue to "generate" new constructs at a rapid pace. All of these newly proposed constructs suggest that marketing is a vital discipline characterized by a lot of creativity on behalf of marketing researchers. A brief presentation of this development can be found in Chapter 2.

The majority of these newly proposed constructs are validated based on the approach[2] suggested by Classical Test Theory (CTT), a dominant paradigm for addressing measurement problems in marketing research. However, due to inherent shortcomings of Classical Test Theory (e.g., test properties are sample dependent; rely on strict assumptions; etc.), reliable and valid measurement should not only rely on CTT. Two alternative measurement theories, Item Response Theory (IRT) and Generalizability Theory (GT), have recently been introduced in the marketing measurement literature (see Table 5 and Table 6 in Appendix). Although IRT and GT have not to the same extent been adopted as CTT, some researchers (e.g., Webb, 1991; Embretson and Reise, 2000) claim that they are superior to CTT. However, other researchers (e.g., Peter, 1981) insist that CTT is better suited for the measurement of marketing constructs. This paper focuses on how these measurement theories (i.e., CTT, IRT, and GT) differ, in particular with regard to construct validation.

To address this issue, I will first introduce the definitions of some key concepts in Chapter 3 and 4 and then address how constructs are validated based on Classical Test Theory model in Chapter 5. Also, I will present the alternative measurement approaches, Item Response Theory (IRT) and Generalizability Theory (GT), in Chapter 5. As compared to Classical Test

[2] A factor analytical framework is typically used to conceptualize the constructs (e.g., Churchill paradigm), but the justification of the validation procedures will typically be found in Classical Test Theory.

Theory (CTT), which focuses on the covariance between true score and observed scores and allows only a random error component, Item Response Theory (IRT) uses a non-linear equation to map an individual's item responses towards a particular underlying construct or attribute, whereas Generalizability theory (GT) acknowledges multiple error components and enables researchers to explicitly address reliability issues with its focus on redesigning studies. Consequently, the measurement results towards the same construct might be different according to the different measurement theories. In Chapter 6 I will contrast the three different measurement theories with a particular focus on how they conceptualize constructs and on the validation procedures. Finally, Chapter 7 contains discussion and the conclusions of the theoretical assessment of the measurement theories.

2. The Introduction of New Constructs in Marketing: A Growing Trend

In the past 50 years we have witnessed a tremendous growth of new marketing constructs. Studies have produced a substantial body of knowledge about consumer choice, attitude, satisfaction judgments, consumption meanings, consumer brand relationships, etc. Scholars frequently borrow and develop theoretical propositions such as latent construct, i.e. phenomena that they believe to exist but that cannot be observed directly, to seek explanations for the behavior of consumers or others. Hence, these latent constructs are central in the development of marketing theories.

In different areas of marketing, especially those closely related to human psychology and behavior, a large number of constructs are defined and measured. Brand loyalty, the commitment to rebuy a preferred brand in the future, is created to measure repetitive same-brand purchasing (Oliver, 1999). Brand equity is defined to measure the assets and liabilities linked to a brand (Aaker, 1991). Brand awareness deals with the likelihood that a brand name will come to mind and the ease with which it does so (Keller, 1993). Brand image is the perceptions about a brand reflected by the brand associations held in consumer memory. Brand personality is defined to reflect human like qualities (personality) associated with the brand and researchers have found that consumers are more likely to purchase a brand if its personality is similar to their own. Table 1 lists some frequently studied and measured constructs in marketing. It should be pointed out that the listed constructs are only examples from a much more extensive list found in the marketing discipline.

Table 1: Selected widely accepted constructs in marketing research

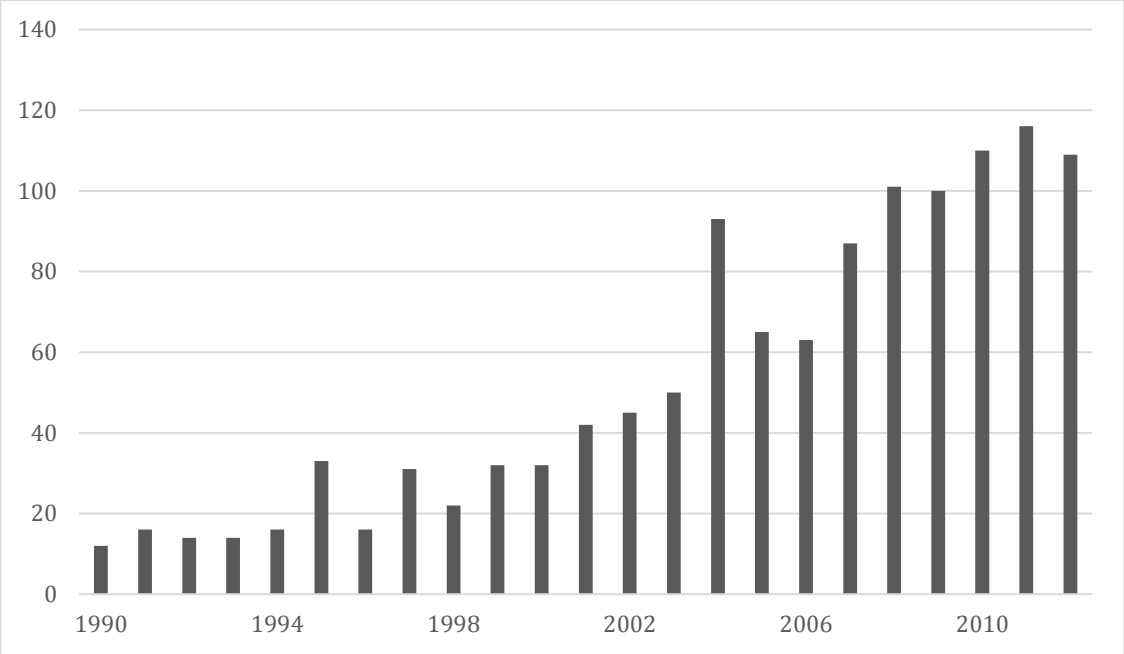
Brand management	Brand loyalty	Day (1976), Jacoby & Kyner (1973)
	Brand equity	Keller (1993)
	Brand awareness	Percy and Rossiter (1992)
	Brand image	Park, Jaworski & Maclnnis (1986)
	Brand personality	Aaker (1997)
	Brand experience	Schmitt, Zarantonello & Brakus (2009)
	Brand Commitment	Morgan & Hunt (1994)

Consumer relationship	Trust	Morgan & Hunt (1994)
	Satisfaction	Churchill & Surprenant (1982)
	Experience	Novak, Hoffman & Yung (2000)
	Service quality	Parasuraman, Zeithaml & Berry (1985)
	Relationship quality	Crosby, Evans & Cowles (1990)
	Word of mouth	Richins (1983)
Societal marketing	Conflict handling	Song, Xie and Dyer (2000)
	Green marketing	Peattie & Charter (1994)
	Benevolence	Selnes & Gønhaug (2000)
	Ethical codes	Singhapakdi & Vitell (1990)
Business marketing	Information exchange	Menon & Varadarajan (1992)
	Collaboration	Gummesson (1996)
	Personalization	Walters & Lancaster (1999)
	Bonds	Brown & Dacin (1997)
	Availability	Chaston & Baker (1998)

Quite a few of these constructs are well-established and widely accepted in academia and among practitioners. Further, the speed of creating new latent constructs in marketing seems to be much faster in recent years than in previous decades. To support this proposition I conducted a targeted investigation of the recent marketing literature. Since marketing constructs often are introduced with different names, definitions or measurements, it is not realistic to capture the number of all recently created constructs in a simple literature search. Hence, the following results will serve just as an approximation. I first conducted a limited search for literature which involved “construct”, “latent variable”, and “factor analysis” in a selected number of marketing journals. Although this search covers both new constructs and existing constructs, the results reveal an interesting pattern (see Figure 11 in Appendix). I find that references to latent construct and the measurement of latent constructs have shown a growing trend from 1991 to 2011. Certainly, this trend by itself does not provide support to the claim that there is a rapid growth in new constructs within marketing, but simultaneously it appears fairly safe to assume that new constructs would be a substantial part of these latent constructs assessed in the marketing literature.

A next step was to search abstracts in the EBSCO business source premier database, subject marketing, using the following key words: "conceptualize", "conceptualization" or "conceptualizing". The key words are determined based on an examination of commonly used terms in a small sample of marketing articles when they refer to procedures used to define new constructs. Although constructs are introduced and analyzed under different frameworks, to create a new latent construct, it is especially common and necessary to "conceptualize" it first. It is believed that this procedure will rarely be found in papers that do not intend to present new latent constructs. Although these key words will not capture all new constructs introduced in the literature, they might provide an indication. The search was conducted for each year from 1990 to 2012 and the results are presented in Figure 2. Figure 2 reveals that the number of papers referring to the conceptualization of constructs is increasing from the late 1990s up to 2011 and hence these results may be taken as an indication of an increasing trend of introducing new concepts in the marketing literature.

Figure 2: Growth of marketing literature that performs "conceptualization" from EBSCO database



This trend can be explained as a response to new requirements facing marketing research, which is asked to deal with human behavior and psychology under different circumstances. Below I present a few speculations regarding what might account for the apparent increase in new marketing constructs.

Hybrid or decomposed construct

Existing constructs are frequently challenged by emerging new concepts that might provide better explanations of market behaviors. Consequently, new latent constructs will be derived from existing constructs when adjustments of the definitions are needed. Those adjustments include both combining overlapping constructs and decomposing existing dimensions. For example, a widely used construct, brand equity, is defined as a combination of other constructs, such as brand awareness and brand associations (Keller 1993). A newly proposed construct, Brand Love (Carroll and Ahuvia, 2006), is in fact an outcome of restructuring several existing constructs such as brand loyalty and brand experience.

Multi-dimensional constructs

Latent constructs are normally hierarchical and multi-dimensional. For instance, "Attribute Satisfaction" is conceptualized as a multidimensional and multilevel construct with three primary dimensions: the core of the service, the peripheral aspects of service quality (SQUAL), and value. Furthermore, SQUAL has three sub-dimensions and value has two (Gonçalves, 2013). Such new dimensions, defined as latent constructs, also contribute to the rapidly growing number of new constructs.

Context based constructs

Researchers also keep creating new constructs derived from existing theoretical constructs based on specific research requirements. For example, internet marketing is particularly designed to adjust the main stream marketing theories with properties of the web consumer communities. Brand prominence is defined to better understand the signaling status with luxury goods. Constructs like green brand image, green brand trust, or green brand association are created to capture the research context of environmental responsibility. In fashion industry, attitudes such as cool or stylish, are defined as extra indicators to measure marketing performance.

Not all new constructs necessarily follow these patterns mentioned above. Constructs can also be created based on metaphorical transfer (Breivik and Thorbjørnsen, 2008) where

constructs are transferred from a different source domain to marketing. For instance, Brand Relationship Quality (Fournier 1998) is taken from the source domain of social psychology, in particular the one of marriage (Breivik and Thorbjørnsen, 2008). Taken together these sources of new constructs suggest that new constructs will continue to pop up in marketing. Hence, construct validation would be expected to be a much focused area. However, it has been suggested that marketing researchers at least prior to 1980 have given little explicit attention to construct validation (e.g., Peter, 1981).

Recent practice in marketing research (after the debate in the beginning of the 1980s) does pay more attention to construct validation. Reporting the estimates of internal consistency reliability (e.g., Cronbach's alpha) and conducting factor analysis have become standard procedures. Particularly, the introduction of structural equation models (SEM) has led to improved procedures for construct validation (Jarvis, MacKenzie and Podsakoff, 2003).

However, latent construct validation is complex. To better understand issues related to construct and construct measurement, I will first review some basic concepts from a measurement perspective. The two following chapters address the definition of concepts and constructs as well as the process of how to measure these constructs.

3. Marketing Construct and Latent Variable

3.1 Construct

3.1.1 Definition

Peter (1981, pp.133) concluded that "a construct is a term specifically designed for a special scientific purpose, generally to organize knowledge and direct research in an attempt to describe or explain some aspect of nature." A construct can be an idea that unites phenomena (e.g., attitudes, behaviors, traits) under a single term. This term can be abstract or concrete (Bollen, 1989). It is created by people who believe that some phenomena have something in common; and it identifies what thing or things do(es) this(these) phenomena have in common.

Kaplan (1964) stated that constructs have at least two types of meaning, systemic and observational. Systemic meaning refers to the interpretation of what a construct stands for depends on the theory in which the construct is embedded. For example, to know what a researcher means when he discusses the construct "loyalty", we must know which theory of "loyalty" the researcher is using. Observational meaning refers to the notion that a construct must be capable of being directly or indirectly operationalized if it is to have explanatory power (Torgerson, 1958). If a construct has no observational meaning, it is merely a metaphysical term; if a notion has no systemic meaning, it is not a construct but an observational term (Peter, 1981).

Constructs range from simple to complex and vary in level of abstraction. Researchers often use the terms construct and variable interchangeably. However, some researchers (Hale and Astolfi, 2011) claim that constructs are latent, that is not observable, and cannot be measured directly, whereas variables can be either observable or unobservable. Furthermore, the constructs of particular interest to marketing would typically be behavioral as opposed to those found in the physical sciences, and as such constructs must be behavior-relevant and should be embodied in at least a loosely specified theory to be of interest to marketing researchers.

3.1.2 Construct and Concept

Hale and Astolfi (2011) suggested that concepts are words or symbols with an implied commonly understood meaning and are commonly accepted as labels for a specific event, situation, behavior, attitude, and so forth. Cooper and Emory (1995) indicated that concepts are created by classifying and categorizing objects or events that have common characteristics beyond the single observation, and can be measured directly or indirectly.

A concept is the prerequisite for and the basis of a construct (see Figure 1). Sometimes concepts are combined to determine a construct. Hence, construct needs to be conceptualized before it can be measured. MacKenzie et al. (2011) suggested that identifying what the construct is intended to measure and how the construct differs from other related constructs is a part of the first stage of scale development and validation process. Given the existence of the concept (as we assumed in this paper), success of research hinges on how clearly we conceptualize the construct and how well others understand the construct(s) we use (Cooper and Emory, 1995). During the conceptualization stage, MacKenzie et al. (2003) pointed out that the researcher should specify the nature of the construct and its content in unambiguous terms and in a manner that is consistent with prior research.

3.2 Latent Variable

The term variable is used by scientists and researchers as a symbol to which numerals or values are assigned. Latent variables represent abstractions that permit us to describe or measure constructs which cannot be directly observed in the model. Latent variables and constructs are both terms that researchers use to refer to abstract objects in scientific studies. Often the terms are used interchangeably. However, latent variables are more commonly referred to when we statistically model our data.

Latent variables have been found so useful that they pervade virtually all fields of social science (Glymour et al., 1987). Obviously, marketing is not an exception. There appears to be no single general definition of a latent variable that could encompass all the diverse applications of latent variables. Bollen (2002) summarized some non-formal definitions of latent variables, such as latent variables are hypothetical variables, they are unobservable or

unmeasurable, and they are used as a data reduction device. Furthermore, he (2002) also pointed out some ways to formally define latent variables. Examples of formal definitions of latent variables can be local independence, expected value true scores, and nondeterministic functions of observed variables.

Under the local independence definition (Hambleton et al., 1991; Bollen, 2002), observable variables are independent to each other if latent variables are held constant. To assess local independence we need at least two observed variables. For instance, if we assume that the construct "consumer loyalty" is responsible for the correlations between the observable variables "last purchase" and "purchase ratio", these two observables should not be correlated when we control for the effect of "consumer loyalty". Hence, latent variables are defined based on their ability to completely explain "away" the association (dependence) found between indicators supposed to measure the latent variable.

The expected value definition of a latent variable is referred to as the "true score", which is measured by the mean of repeated observations of a variable for an individual. As an illustration, if we use "last purchase" as an indicator of "consumer loyalty" for an individual, then the mean value of "last purchase" for a repeated set of observations is the true score for this individual under the hypothetical situation that each repeated observation is independent. This is the basis for Classical Test Theory (CTT). True score would be the mean of the repeated observations of variable values for an individual[3].

Latent variable is defined by Bentler (1982, p.106) as a nondeterministic function of observed variables: "A variable in a linear structural equation system is a latent variable if the equations cannot be manipulated so as to express the variable as a function of manifest variables only". In the "consumer loyalty" example, "consumer loyalty" is a latent variable if it cannot be exactly expressed by its measures, say "last purchase" or "purchase ratio". This definition makes it clear that we cannot use observed variables to exactly determine the latent variable. We might be able to estimate a value on the latent variable, but we would not be able to make an exact prediction just based on its observed indicators.

[3] Since it is almost impossible to imagine truly independent repeated observations in the social science, the expected value must be based on non-observable counterfactuals.

Bollen (2002) proposed an alternative definition for latent variables: "a latent random (nonrandom) variable is a random (nonrandom) variable for which there is no sample realization for at least some observations in a given sample". All variables are latent until sample values of them are available (Bollen, 2002). This definition permits a random variable to be latent for some cases but manifest for others, since a variable might have values in some cases and might not have values in some other cases. Hence, previously latent variables might be observable when we develop better measurement instruments.

"Sample realization" definition is more simple and inclusive than the other definitions presented above (Bollen, 2002). For example, both latent variables as defined by local independence and the expected value definition would be special cases of the sample realization definition, while some variables would qualify as latent based on the local independence definition, but not with regard to the expected value definition and vice versa. This could lead to counterintuitive elimination of some variables as latent variables. For instance, the expected value definition associated with Classical Test Theory, assumes there is a linear relationship between the latent variable and its indicators, since the value of the latent variable is the mean of observed variables values. In contrast, Item Response Theory suggests that there is a nonlinear function connecting the items and the underlying latent variable[4]. Thus, in IRT models, the underlying variable (i.e., the trait level θ) does not qualify as a latent variable according to the expected value definition.

Given the purpose of studying different measurement theories, an inclusive definition of latent variables is required. Thus, the more general definition offered by Bollen (2002) is deemed most relevant for the present study.

[4] The nonlinear relationship between items and the underlying latent variable found in Item Response Theory will be discussed later in Chapter 5.3.

4. Measurement

4.1 Definitions

"Whatever exists at all exists in some amount" (Thorndike, 1918, p.16). Accordingly, if we believe or assume some marketing constructs exist, these constructs should exist in some quantity. Measurement is used to determine the quantity.

There are many definitions of measurement. In 1946, Stevens provided a straightforward definition of measurement as can be seen from the following quote: "the assignment of numerals to objects or events according to rules" (Stevens, 1946, pp.677). Nunnally (1967) suggested that the process of measurement or operationalization involves rules for assigning numbers to objects to represent quantities of attributes. Bollen (1989) also defines measurement as a process by which a construct is linked to one or more latent variables and how these are linked to observed variables. Another term employed to characterize the assignment of numbers to measure behavioral and psychological constructs is scaling, because the assignment of numerals places the objects or events on a scale (Jones and Thissen, 2007). To make use of these data, the "rules" for the assignment of numerals are usually based on the measurement theory. Specifically, a measurement theory specifies correspondence rules for linking empirical observations (observables) to abstract latent variables[5] (Blalock, 1968; Weiss and Davison, 1981).

4.2 Scaling

Scaling evolved out of efforts in psychology and education to measure "unmeasurable" constructs such as self-esteem (Trochim, 2000). For some researchers, the terms scaling and measurement are synonymous (e.g., Bartholomew et al., 1996; Jones and Thissen, 2007). Other researchers reserve the term scaling to the assignment of numbers that, at a minimum, have the property of order (McDonald, 1999). Still more restrictive definitions require the use of scalars (Wright, 1997). However, in this paper, I employ a broader definition of scaling,

[5] Three major measurement theories will be discussed in Chapter 5.

which refers to the specific way that numbers or symbols are linked to behavioral observations to create a measure (Allen and Yen, 1979; Crocker and Algina, 1986).

4.2.1 Purpose of Scaling

In most scaling tasks, the objects are text statements, such as statements of beliefs or attitudes. Through scaling, researchers would assign numbers or symbols to the participants' responses towards these statements (Trochim, 2000). For instance, if there are several statements towards aggressiveness, and if a participant provides his or her responses to each statement, then researchers would infer the person's overall attitude towards aggressiveness via scaling. Scaling can help to test hypotheses (Trochim, 2000). For example, scaling enables researchers to test whether a construct is unidimensional or multidimensional. With several questions or items measuring a concept, we would use scaling to figure out what dimensions might underlie a set of ratings, and determine how well these items are connected and hence examine whether they measure one concept or multiple concepts.

4.2.2 Scale

Measurement instruments used to combine sets of items into a composite score are often referred to as a scale (DeVellis, 2003). The scale score is intended to provide levels to latent variables not readily observable by direct means. We develop scales to measure phenomena that we believe to exist[6], but that we cannot assess directly. Although researchers are interested in values of constructs rather than items or scales, measuring a construct is normally based on a measurement model of the correspondence between observable items and the underlying construct of interest. For example, a market researcher might be interested in assessing customers' commitment towards a product. However, since this cannot be directly observed she might try to infer the level of commitment based on the customer's responses to a set of questions tapping different aspects of commitment. The observed scores from the questionnaires are then transformed to a scale score based on some sort of measurement model. The derived scale scores would then be used as measures of the latent construct.

[6] Typically based on our theoretical conceptualization of the "real world" phenomena we want to understand (the conceptual model)

Researchers can use multiple items or a single item to measure a marketing construct. However, in social science researchers typically would claim that no single item is likely to provide a perfect representation of a construct, in that no single word can be used to test for differences in subjects' spelling abilities and no single question can measure a person's intelligence. Churchill (1979) further pointed out that, single items, which potentially contain considerable measurement error, tend to have only a low correlation with the construct being measured and tend to relate to other attributes as well. Thus, respondents usually respond to two or more measures intended to be alternative indicators of the same underlying construct (Gerbing and Anderson, 1988).

4.2.3 Unidimensional Scale

Multiple items or measures might be indicators of different dimensions of a construct. For instance, the "score" on a bathroom scale does reflect one and only one dimension, which is the weight; while the measures of intelligence normally will reflect more than one dimension, such as logical and verbal ability as well as other aspects (i.e., previous experience).

If the scale only reflects one dimension, it will be referred to as unidimensional. Unidimensionality is a critical aspect of measurement theory. The measure would be valid only if related items measure only one underlying construct (Hattie, 1985).

Thurstone (1931) came up with three different methods for developing unidimensional scales: the method of equal-appearing intervals; the method of successive intervals; and the method of paired comparisons. All of them begin by focusing on a concept that is assumed to be unidimensional. They also involve starting out with generating a large set of potential scale items or statements with the end of constructing a final scale consisting of relatively few items which the respondents rates on an agree or disagree basis (Trochim, 2000). The major differences among the methods refer to how the data from the judges is collected. For instance, the equal-appearing intervals compute the scale score value (e.g. median, interquartile range) for each item and hence require the respondent to attend to each item sequentially while the paired comparisons require each judge to make judgments about each pair of statements, a time-consuming process if there are numerous items or statements involved.

There are also other scales that assume unidimensionality, such as Likert scaling and Guttman scaling. Likert scaling utilizes 1 – to – 5 or 1 – to – 7 response scales according to respondents' own degree of agreement or disagreement, and selects the items from all the candidate items through computing the intercorrelation between all pairs of them, based on the ratings of judges (Likert, 1931). Guttman scaling, which is also known as cumulative scaling, aims to establish a one-dimensional continuum for a concept of interest (Torgerson, 1962). Items are arranged in an order so that an individual who agrees with a particular item also agrees with items of lower rank-order. In particular, Guttman scaling constructs a matrix or table that shows the responses of all respondents on all the items, sorts this matrix, and attempts to identify a cumulative scale by the scalogram analysis.

4.2.4 Multidimensional Scale

If a scale refers to more than one dimension, it is referred to as multidimensional. The intelligence scale often postulates two major dimensions – mathematical and verbal ability. Some participants would be high in verbal ability but low in mathematical ability, or vice versa. In this case, the scale of intelligence is multidimensional. Accordingly, it is not accurate to depict a person's intelligence assuming a unidimensional scale.

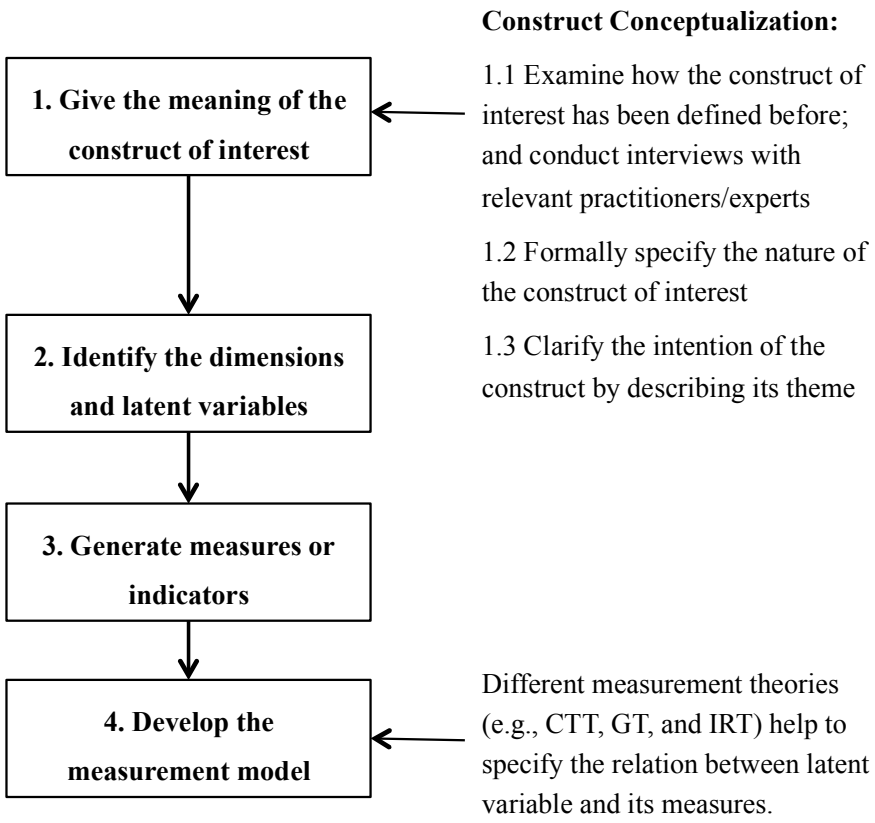
Various strategies for multidimensional scaling (MDS) borrow from Thurstone's scaling models (1958) and their descendent, the idea that similarity data (as might be obtained using various experimental procedures and human judgment) can be represented spatially (Jones and Thissen, 2007). But Thurstone's scaling models represent objects with real numbers on a single dimension, while multidimensional scaling represents objects as points in two- (or higher) dimensional space. Moreover, multidimensional scaling is a collection of techniques that represent proximity data in such a way that those data corresponding to similar stimuli are located together, while those corresponding to dissimilar stimuli are located far apart (Jones and Thissen, 2007).

4.3 Measurement Process

Bollen (1989) proposed a stepwise procedure to measurement involving four stages (see Figure 3). The first choice, however, is to decide on which constructs one wants to examine

and hence measure. Once a construct is selected the measurement process would follow the four steps (Bollen, 1989). The first step is to give the meaning to the construct. Researchers should develop a theoretical definition, which explains in as simple and precise terms as possible the meaning of a construct. Usually it helps to link an abstract construct with some conceptual terms, and distinguish exactly what kind of relationships we expect to find between our focal construct and other related constructs[7]. From the theoretical definition, dimensions, the distinct aspects of a construct, can be identified.

Figure 3: A four-step measurement process



Mackenzie et al. (2011) outlined a three stage process for the construct conceptualization step. The first stage is that researchers need to examine how the construct of interest has been defined in prior research, and to conduct interviews with relevant practitioners and/or experts. Next, "researchers need to formally specify the nature of the construct, including (1) the conceptual domain[8] to which the construct belongs and (2) the entity[9] to which it applies"

[7] Referred to as nomological net.
[8] The conceptual domain represents the phenomena to which the construct refers (i.e., the intension).
[9] The entity represents the referents to which the construct applies (i.e., the extension).

(Mackenzie et al., 2011, p.259). By defining conceptual domain, researchers should specify the general type of property to which the construct refers. For example, the definition should specify whether the construct refers to a thought (e.g., cognition, value, intention), a feeling (e.g., attitude, emotion), an outcome (e.g., performance, return-on-investment), an intrinsic characteristic, etc. By defining the entity, researchers should know the object to which the property applies (e.g., a person, a task, a process, a relationship, an organization, a culture and etc.). For instance, according to Doll and Torkzadeh (1988) the definition of "end-user computer satisfaction" should focus on a person's (entity) positive feelings about computer technology (general property). According to Davis (1989) the definition of "perceived ease of use of technology" should focus on a person's (entity) perception regarding the use of information technology (general property).

Finally, researchers need to clarify the intension of the construct by describing its conceptual theme. The conceptual theme of a construct consists of the set of fundamental attributes or characteristics that are necessary and sufficient for something to be an exemplar of the construct. It is important that this conceptualization is conducted in a clear and concise language to avoid multiple interpretations and overly technical descriptions.

The second step in the measurement process requires researchers to identify the dimensions and corresponding latent variables for the construct of interest. As many constructs have numerous possible dimensions, a definition is critical to set the limit on the dimensions a researcher wants to investigate. Normally, one latent variable represents one dimension.

The third step is to generate measures or indicators. This step is sometimes referred to as operationalization, which describes the procedures for forming measures of the latent variable(s) that represent a construct. For instance, an operational definition could be a survey questionnaire, a method of observing events in a field setting, a way to measure symbolic content in the mass media, etc. An operational definition or measure is appropriate to the extent that it leads to an observed variable that corresponds to the meaning assigned to a construct. Then researchers would measure these observed variables empirically. For example, in some situations latent variables are operationalized as the responses to questionnaire items, census figures, or some other observable characteristics.

The formulation of these measures is also guided by the theoretical definition. Specifically, the theoretical definition helps researchers to ensure whether a phenomenon or an observable variable is encompassed or excluded by the construct of interest, and thus helps to measure the corresponding content of the construct. Therefore, a theoretical definition serves several important and useful functions. It links a term to a specific construct, identifies its dimensions and the number of latent variables, and sets a standard by which to select measures.

The last step in the measurement process is to specify the relationship between the latent variable and its measures. This relation can be constructed based on different theories. For example, Classical Test Theory and Generalizability Theory imply a linear relation between the latent variable and its measures, while Item Response Theory suggests the existence of a nonlinear relation.

5. Measurement Theories

5.1 Classical Test Theory (CTT)

5.1.1 Theory Overview and Definition

Classical Test Theory (CTT), regarded as roughly synonymous with True Score Theory, is the most often used theory in psychological testing (Borsboom, 2009), and is widely utilized in social sciences. Nowadays, considerable research towards latent variables still predominantly follows CTT (Salzberger & Koller, 2012). The axiomatic system of this theory was introduced by Novick (1966), and formed the basis of the most articulate exposition of the theory (Lord and Novick, 1968).

The fundamental idea of Classical Test Theory is that an observed score is the result of the respondents' true score plus error: $X_O = X_T + X_E$. Hence, the error score, X_E, is defined as the difference between the observed score and the true score. From the equation, we can identify that there is a linear relation between the true score and the observed score. Thus, Classical Test Theory provides a simple way to link the latent variable[10] (i.e., true score) with its manifest variables (i.e., observed score). Furthermore, Classical Test Theory defines the true score of person i towards a measurement, X_{Ti}, as the expectation of the observed score X_{oi} over replications: $X_{Ti} \equiv \sum(X_{Oi})$.

The True Score

In Classical Test Theory, the true score is the expectation of the observed score X_{oi} over replications. However, it is not possible to obtain an infinite number of replications (i.e., test scores), so X_T is hypothetical. Furthermore, in general, the true score does not admit a realist interpretation (Borsboom, 2009). The true score is syntactically defined in terms of a series of observations, therefore it cannot exist independently of its observed scores. The true score

[10] As discussed in Chapter 3, Classical Test Theory uses "true score" to represent the underlying latent variable.

can only apply to the test in terms of which it is defined, and thus has a highly restricted domain of generalization. Accordingly, Borsboom (2009) suggested that true scores should be conceptualized as an instrumental[11] concept that governs the interpretation of data analytic results in test analysis, rather than an entity that exists independently of the researcher's imagination.

True score is also commonly introduced by using phrases such as "the true score is the construct we are attempting to measure" (Judd, Smith, and Kidder, 1991, p.49), or by stressing the distinction "between observed scores and construct scores"(Schmidt and Hunter, 1999, p.189). This interpretation of true scores can be referred to as the platonic true score (Lord and Novick, 1968). The "platonic true score" suggests that the true score accurately reflects a latent variable. However, many researchers (Klein and Cleary, 1967; Lord and Novick, 1968) pointed out that this interpretation is sometimes untenable and would cause some problems. For example, in some cases equating the true score with the construct score leads to correlations between true and error scores (see Klein and Cleary, 1967), which in turn violates the assumptions of CTT.

The Error Score

Classical Test Theory assumes that the error term, X_E, is random. The random error, considered as noise, is caused by any factors that randomly affect measurement of the variable across the sample (Trochim, 2000). For instance, the errors might exist due to some transient personal factors, situational factors (e.g. whether the interview is conducted in the home or at a central facility), mechanical factors (e.g. a check mark in the wrong box or a response which is coded incorrectly), etc. The important property of random error is that it adds variability to the data but does not affect average performance for the group (Trochim, 2000). The primary objective in CTT measurement is to produce X_o which approximate X_T as closely as possible (Churchill, 1979). Thus, it is to reduce the inconsistency caused by the measurement error.

[11] In the instrumentalism, usefulness is the only appropriate criterion, and whether a theoretical attribute exist or not is unimportant.

5.1.2 Basic Classical Test Theory Assumptions

The mathematical equation of Classical Test Theory illustrates the first simple but fundamental theoretical assumption of CTT, that observed scores on a marketing construct's measures are determined by respondents' true scores and by measurement error scores (Furr and Bacharach, 2008). Unfortunately, the researcher never knows for sure what the X_T scores and X_E scores are. Rather, the measures are always based on inferences, and the quality of these inferences depends directly on the procedures that are used to develop the measures and the evidence supporting their "goodness" (Churchill, 1979). This evidence typically takes the form of some reliability and validity index (Churchill, 1979).

In addition, CTT also rests on the following assumptions: (1) the items are randomly sampled from a larger item domain (Furr and Bacharach, 2008), (2) measurement error occurs as if it is random, and (3) the inflation and deflation caused by measurement error is independent of the individual's true levels of the underlying variable. An important consequence follows the randomness assumption is that errors tend to cancel themselves out across respondents. Thus, the expected value of the error should be zero.

Assumptions of Classical Test Theory in the Reality

Classical Test Theory requires the replications to be parallel[12] in order to keep the true score invariable over time. However, as CTT only assumes measurement on a single occasion rather than a series of measurements, replications of the measurements can hardly be parallel. It is unrealistic for the true score to remain constant over replications, because participants would remember their previous response. They will learn and even change in many other ways during a series of repeated administrations of the same test. Hence it would be problematic to view CTT as concerned with series of measurements. In order to change this awkward situation, Lord and Novick (1968) introduced a brainwashing thought experiment. They assumed that the brainwashing would render the independent replications by "deleting" subjects' memories, and thus enable us to ask the same question over and over again. This is certainly a hypothetical trick for CTT to make sense at least at a hypothetical level. It is of

[12] "Parallel replications" mean that the expected value and error variance for each replication should be identical.

course not relevant for describing how one would expect the model to work on a practical level. However, instead of asking the same question over and over again, it seems more practical to ask different questions but with the identical expected value and error variance (i.e., parallel forms). Hence in reality, researchers present the participant with different questions that are assumed to be parallel, although they all know that finding the truly parallel forms is hard and even unlikely.

5.1.3 Reliability

Features of Reliability

The typical application of Classical Test Theory does not involve testing the model assumptions, but contains the estimation of reliability (Furr and Bacharach, 2008). Generally speaking, the aim of Classical Test Theory is to understand and improve the reliability of measurement tests.

In research, the term reliability means "repeatability" or "consistency". A measure is reliable to the extent that independent but comparable measures of the same construct of a given object agree (Churchill, 1979); or that it consists of reliable items that share a common latent variable. The value of a reliability index of a test instrument depends on the specific sample. For example, if two subsamples of students with different levels of mathematical abilities (moderate and low) take the same mathematical ability test, the measurement results of the students with moderate mathematical ability tend to be more consistent than the students with low mathematical ability. Thus, reliability can be meaningfully considered only when interpreted in terms of individual differences in a specific population (Borsboom, 2009). Hence, reliability confounds measurement quality and the sample characteristics.

I will like to address two other aspects of reliability. First, reliability is not an all-or-none property of the results of a measurement procedure but is located on a continuum (Furr and Bacharach, 2008). A procedure for measuring something can be more or less reliable. Second, it is itself a theoretical notion (Furr and Bacharach, 2008). Just as satisfaction or self-esteem is an unobserved feature of a person, reliability is an unobserved feature of test scores. It cannot exist outside of its test sample and its test scores.

Reliability: Classical Test Theory Perspective

According to Classical Test Theory, reliability is a test property that derives its meaning from observed scores, true scores, and measurement error. It reflects the extent to which differences in respondents' observed scores are consistent with differences in their true scores. More specifically, the reliability for a measurement procedure depends on the extent to which differences in respondents' observed scores on the measure can be attributed to differences in their true scores, as opposed to other, often unknown, test and test administration characteristics (Furr and Bacharach, 2008). The extent to which these "other" characteristics contribute to differences in observed scores is referred to as measurement error, because they create inconsistency between observed scores and true scores.

Reliability also hinges on the links among observed score variability, true score variability, and error score variability[13]. The variance of a composite score is determined by the variability of each item within the composite, along with the correlations among the items. Consequently, the relation among observed variance, true score variance and error score variance can be seen in the following equation: $\sigma_O^2 = \sigma_T^2 + \sigma_E^2 + 2r_{TE}\sigma_T\sigma_E$. Total observed score variance should be equal to true score variance plus error variance plus the covariance of true scores and error scores. However, since the error is independent of the true score, the correlation between error score and true score is zero. Therefore total variance of the observed scores from a group of respondents will equal the sum of their true score and error score variances: $\sigma_O^2 = \sigma_T^2 + \sigma_E^2$.

In Classical Test Theory, there are at least four ways that can be used to define reliability (Furr and Bacharach, 2008). Each of these conceptual approaches arises from the associations among observed scores, true scores, and measurement error. The most commonly used definition is the squared population correlation, ρ_{XT}^2, between true and observed scores: $\rho_{XT}^2 = \frac{\sigma_T^2}{\sigma_O^2} = \frac{\sigma_T^2}{\sigma_T^2+\sigma_E^2}$. Accordingly, the value of the reliability coefficient will decrease as the error variance increases. If there is no error variance, reliability is perfect and equals unity. Similarly, if the true score variance in a population approaches zero while the error variance remains constant, the reliability becomes smaller.

[13] The error score variability here assumes to be fully random.

Table 2: A 2x2 framework for conceptualizing reliability (Furr and Bacharach, 2008)

		Conceptual Basis of Reliability: Observed Scores in Relation to	
		True Scores	Measurement Error
Statistical Basis of Reliability in Terms of	Proportions of Variance Correlations	Reliability is the ratio of true score variance to observed score variance $\rho_{XT}^2 = \frac{\sigma_T^2}{\sigma_O^2}$	Reliability is the lack of error variance $\rho_{XT}^2 = 1 - \frac{\sigma_E^2}{\sigma_O^2}$
		Reliability is the (squared) correlation between observed scores and true scores $\rho_{XT}^2 = r_{OT}^2$	Reliability is the lack of correlation between observed scores and error scores $\rho_{XT}^2 = 1 - r_{ET}^2$

Estimates of Reliability

There are three main ways to empirically estimate reliability and hence reliability can be divided into three different types based on its estimation methods. The first method is known as the test-retest method, which assumes the interpretation of actual repeated measurements as identical[14]. It involves having the same set of people complete two separate versions of a scale or the same version on multiple occasions.

The second method is based on the idea that two distinct tests could be parallel. Researchers use the correlation between the observed scores on two identical (parallel) administrations (i.e., tests) to determine the reliability. The third method suggests that reliability can be computed based on the covariance of the parallel items within a test (e.g., internal consistency). The rationale underlying this reliability estimation is that each item of a scale is exactly as good a measure of the latent variable as any other of the scale items. If the individual items are strictly parallel, the relationships among items can be logically connected to the relationships of items to the latent variable.

[14] To meet this assumption Lord and Novick (1968) developed the brainwash example. This makes the rationale hypothetical and it must be safe to assume that this assumption never is met in social sciences.

Test-retest reliability

The test–retest method is based on the idea that two administrations of the same test may be regarded as one administration of two parallel tests (Borsboom, 2009). If this were the case, the population correlation between the scores on these administrations would be equal to the reliability of the test scores. This approach assumes that there is no substantial change in the construct being measured between the two occasions. In other words, the latent variable (i.e., the true score) should be stable to make the repeated administrations parallel.

However, many researchers have questioned whether the test-retest index can reliably measure "reliability". Borsboom (2009) defined the test-retest correlation as "stability coefficient", and pointed out that between-subjects correlations cannot distinguish between situations where individual true scores are stable and situations where they increase or decrease by the same amount. Thus the stability coefficient should only be taken to refer to the stability of the ordering of persons, not to the stability of the construct itself. McDonald (1999) also pointed out that to the extent the test-retest method confounds unreliability and differential change trajectories, which might be homogeneous or heterogeneous across subjects, we would find it problematic treating it as if it is a reliability estimate.

In addition, the test-retest method does not provide guidelines as to how to choose an appropriate spacing of the replications. As the correlation between the two observations will depend in part by how much time elapses between the two measurement occasions, it is crucial to decide the amount of time allowed during the replications (Trochim, 2000). The shorter the time gap, the higher the correlation; the longer the time gap, the lower the correlation. Therefore if the test-retest correlation is treated as the only estimate of reliability, researchers might obtain considerably different estimates depending on the time interval or spacing of the measurements.

Parallel-forms reliability

As it is impossible for researchers to determine the true score variance, they would use the correlation between the observed scores on two identical administrations, X and X', to

determine the reliability: $\rho_{XX'} = \frac{\sigma_{XX'}}{\sigma_O^2} = \frac{\sum(TT')}{\sigma_X + \sigma_{X'}} = \frac{\sigma_T^2}{\sigma_O^2}$. As long as the two variables, X and X', are one and the same (i.e. $T \equiv T'$), the possible connection would be created, because equating the correlation between parallel tests with the reliability of a single test makes sense only if the two tests measure the same true score.

To create two parallel forms one might create a large set of questions that address the same construct and then randomly divide the questions into two sets. Then the researchers administer both instruments to the same sample of people. This method assumes that a simultaneous administration of two different tests could be viewed as approximating two replications of a single test. The correlation between them could then be taken to be a direct estimate of the reliability of the test scores.

One major practical problem with this approach is that researchers have to be able to generate lots of items that reflect the same construct. The search for parallel test forms has been unsuccessful, because the empirical requirements for parallelism (i.e. equal means, variances, and covariance of observed scores) are rather demanding. Even by chance the assumption that the randomly divided halves are parallel or equivalent will often not be the case. Furthermore, Borsboom (2009) pointed out that the idea of two distinct tests being parallel seems to be hard to grasp semantically. As the latent variable (i.e., true score) cannot exist independently of its observed scores[15], the true score of T is explicitly defined in terms of the test T, and the true score of T' is semantically interpreted in terms of the test T'. In order to be parallel tests, one test T should be identical to another test T'. Although this can be done by the brainwashing experiment, the true scores on distinct tests T and T' are still semantically distinguishable, simply because they are defined with respect to different tests. The tests might be empirically equal, but this does not make them logically identical.

The parallel forms approach resembles the split-half reliability. The major difference is that parallel forms are constructed so that the two forms are used independent of each other and considered equivalent measures, whereas in a split-half estimation, researchers have just one scale or form as a single measurement instrument and only develop split halves randomly to estimate reliability.

[15] See previous discussion of true score.

Internal Consistency Reliability

The internal consistency reliability is used to assess the consistency of results across items within a test. Researchers use a single measurement instrument administered to a group of people on one occasion. The reliability of this instrument is then judged by estimating how well the items that reflect the same construct yield similar results (Trochim, 2000), or how consistent the results are for different items reflecting the same construct within the measure. A wide variety of measures can be used, such as the average inter-item correlation, average item-total correlation, split-half reliability, and Cronbach's alpha.

The average inter-item correlation uses all of the items designed to measure the same construct in the instrument and computes the correlation between each pair of items. The average inter-item correlation is simply the average or mean of all these correlations. In contrast, the average inter-total correlation computes not only the inter-item correlation but also a total score for the whole set of items and then uses the calculated total score as a new variable in the analysis.

The split-halves method splits a test in two subtests of equal length, assuming the subtests are parallel (or constructs them to be nearly so; see e.g., Mellenbergh, 1994), computes the correlation between the total scores on subtests, and yields an estimate of the reliability of total test scores by using the Spearman-Brown prediction formula[16].

Internal consistency formulas such as the KR_{20} and Cronbach's alpha extend this method. The Cronbach's alpha, which is widely used as a measure of reliability (Churchill, 1979), can be interpreted as the average of all possible split-halves correlations (Trochim, 2000). If the split-halves are parallel, the resulting quantity yields an exact estimate of the reliability of the total test scores. Since parallelism is as troublesome for split-halves as it is for full test forms, these methods fail for the same reasons as the parallel test method (Borsboom, 2009).

[16] Spearman-Brown prediction formula relates reliability to test length and predicts the reliability of a test after changing the test length.

Summary

All these reliability indices are based on inferences made from the observed empirical relations (involving only observables) to theoretical relations (involving observables and unobservable) (Borsboom, 2009). These inferences are normally and widely used to judge whether the measures or the items reflecting the same construct is good or reliable. As various methods can be used to compute reliability, we may need to choose suitable ones in particular situations. For example, if one does not have access to parallel versions of an instrument, using the internal consistency reliability seems more plausible.

However, exactly estimating reliability from observed data is impractical and theoretically questionable. This has prompted CTT to look at worst-case scenarios, and to search for lower bounds for reliability (Guttman, 1945; Jackson and Agunwamba, 1977). Identifying the lower bounds is probably the most viable defense that could be given for the standard practice in test analysis. No matter how bad things may be, researchers try to compute a reliability which will always be higher in the population and hence the lower bound estimation represents a conservative research strategy (Borsboom, 2009).

5.1.4 Validity

Validity is another important aspect of measurement. Researchers define the concept and scope of validity in a variety of ways. Kelley (1927) stated a test is valid if it measures what it purports to measure. Churchill (1979) suggested that a measure is valid when the differences in observed scores reflect true differences on the characteristic one is attempting to measure and nothing else, that is $X_O = X_r$. Mehrens and Lehmann (1984) believed validity can be best defined as the extent to which certain inferences can be made from test scores or other measurements. Itsuokor (1986) defined validity as the degree to which an observational tool provides for objective appraisal of that is observed. Denga (2003) suggested validity is the extent to which a test is truthful, accurate or relevant in measuring a trait it is supposed to measure. Borsboom (2009) proposed the conception of validity concerns the question whether the attribute to be measured causally produces variations in the measurement outcomes.

Thus, validity is concerned with the extent to which the latent variable (construct) is the underlying cause of observed covariation among items and tells us how well the construct's theoretical and operational definitions mesh with one another. It addresses how well an empirical indicator and the conceptual definition of the construct that the indicator is supposed to measure "fit" together. The better the fit, the higher is the measurement validity, whereas poor measurement validity suggests the test does not measure what it is supposed to do. Sometimes, when the error component is not completely random, the covariance under the set of items would be influenced by other systematic factors (e.g., some situational relative stable characteristics). Hence, in reality it seems more reasonable to assume X_E possesses two components, a random error component, X_r , and a systematic error component[17], X_s. The systematic error component, considered as bias, systematically affects measurement of the variable across the sample. Although it violates the assumption of CTT, it also extends CTT by laying out the foundation for another measurement theory, Generalizability Theory.

[17] There might of course be several systematic error components.

Churchill (1979) supposed that if a measure is valid, it is reliable. Reliability is a necessary but not a sufficient condition for validity. Similar as for reliability, validity is also an unobserved feature of test scores. However, validity is more difficult to achieve than reliability. No one could have absolute confidence about validity, but some measures are more valid than others. The reason is that constructs are abstract ideas, whereas indicators refer to concrete observations. There is a gap between the mental pictures about the world and the specific things researchers do to tap this at particular times and places. Although validity cannot be proved, researchers can develop strong support for it. Validity is part of a dynamic process that accumulates support for good measurement over time, and without it, measurement becomes meaningless (Neuman, 2007).

Types of Validity

There are many different forms of validity found in the literature. I will focus on the following four types of validity; face validity, content validity, criterion validity, and construct validity. Each attempts to show whether a measure corresponds to a construct, though their means of doing so differ (Bollen, 1989).

Face Validity

Face validity addresses the question: "on the face of it, do people believe that definition and method of measurement fit" (Neuman, 2007). Weiner and Craighead (2010) formally defined this validity as the degree to which test respondents view the content of a test and its items as relevant to the context in which the test is being administered. Distinct from more technical types of validity, face validity is concerned with the appropriateness, sensibility, or relevance of the test.

Many researchers (e.g., DeVellis, 2003; Downing, 2006) suggested face validity is the least important of the validity types and also the easiest type to accommodate. For instance, an instrument intended to assess the degree to which people answer untruthfully would hardly benefit from having its purpose apparent to respondent. Furthermore, an item that looks as it measures one variable to some test-takers might look like it measures some other construct to another, equally qualified test-takers. Apart from the obvious interpretation that you have a

problem with the measurement, this discrepancy does not provide any guidelines in terms of what to do. Although many criticize face validity, research on technical (empirical) validity (e.g., criterion validity) has shown a significant positive correspondence between face validity and test item accuracy (Holden and Jackson, 1979). Test items demonstrating face validity tend to be found more valid or accurate as compared to items not possessing face validity also according to more technical (empirical) test of validity. Due to the hot debate regarding the usefulness of face validity, a further investigation is still needed.

Content Validity

Content validity concerns item sampling adequacy, that is, the extent to which a specific set of items reflects a content domain (DeVellis, 2003). It is a qualitative type of validity where the domain of a construct is made clear and the researcher judges whether the measures fully represent the domain (Bollen, 1989). A key question for content validity is how do researchers know the construct's domain? To answer this question, they must return to the first step in the measurement process to identify a theoretical definition. The theoretical definition should make it clear the number of and what dimensions the concept is supposed to include. For content validity each dimension of a construct should have one or more measures (Bollen, 1989). A nonrepresentative sample of measures can distort the understanding of a construct and hence it will lack content validity.

However, Bollen (1989) pointed that the major limitation of content validity stems from its dependence on the theoretical definition. For most constructs in the social sciences, no consensus exists regarding the theoretical definition. Thus, the content domain is ambiguous. Researchers need to not only provide a theoretical definition accepted by their peers but also select indicators that fully cover its domain and dimensions. In sum, content validity is a qualitative means of ensuring that indicators tap the meaning of a construct as defined by the researcher.

Criterion Validity

Criterion validity is the degree of correspondence between a measure and a criterion variable, usually measured by their covariation or correlation (Trochim, 2000). Researchers use the

covariation between a standard or criterion and the measure of the construct of interest to assess the quality of the measurement. The validity of an indicator is verified by comparing it with another measure of the same construct in which a researcher has confidence.

There are two forms of criterion used to assess this form of validity; concurrent validity and predictive validity. Concurrent validity involves a criterion that exists at the same time as the measure, whereas predictive validity involves a criterion that occurs in the future (Bollen, 1989). The absolute value of the correlation between a measure and a criterion sometimes is referred to as the validity coefficient (Lord and Novick, 1968). However, obtaining a criterion that directly and exactly matches the latent variable is perhaps theoretically possible, but would in most situations turn out to be practically impossible. Hence, for many concepts in social science, such criteria are not feasible. Furthermore, it is not always clear what to do when one does not obtain a satisfactory correlation between the criterion and the measure of interest. What should be blamed for the lack of correspondence, poor measurement of the construct of interest or poor measurement of the criterion?

Construct Validity

As mentioned earlier, many concepts in the social science are not clearly defined and consequently face validity and content validity may be found difficult to apply. Furthermore, appropriate criteria for some measures may be hard to come by preventing the computation of criterion validity coefficients. Thus, in these situations construct validity may be the only type of validity that realistically can be assessed.

The term “construct validity” refers to the correspondence between a construct which is at an unobservable and a measure of it which is at an operational level (Peter, 1981). The measure should preferably assess both the magnitude and direction of (1) all of the characteristics, and (2) only these characteristics of the construct it is purported to assess. A less precise (but more realistic) definition of construct validity is that it is the degree to which a measure assesses the construct. In this case a measure is a valid representation of the construct to the degree that it assesses the construct and to the degree that the measure is not contaminated with elements from the domain of other constructs or error. Basically, the construct validity of a measure is inferred if the measure behaves as expected according to what substantive

(and psychometric) theory postulates it should behave. For example, if a construct were hypothesized to have three dimensions, a factor analysis of the set of items proposed to measure that construct should result in three meaningful factors that could be interpreted as supportive evidence of construct validity.

Construct validity has two subtypes, convergent and discriminant validity. Convergent validity is relevant when multiple indicators are used to measure a construct. Convergent refers to the degree that items used to measure a construct are associated with one another. Multiple measures of the same construct should operate in similar ways and this is assessed through the similarity between measures measuring the same theoretical construct. Conversely, discriminant validity refers to the degree that measures of different constructs should be less associated with each other than with the underlying constructs they are supposed to measure. For example, if two constructs A and B are independent, measures of A and B should not be associated. The evidence of discriminant validity is the absence of correlation between measures of unrelated constructs.

Construct Validity Estimation

Campbell and Fiske (1959) devised a procedure called the multitrait-multimethod matrix that is useful for estimating construct validity (see Figure 4). The procedure involves measuring more than one construct by means of more than one method (e.g., a paper-and-pencil test, a direct observation, a performance measure), so that one obtains a "fully crossed" method-by-measure matrix. Importantly, constructs should be uncorrelated with the methods. The multitrait-multimethod is a very demanding and to some extent restrictive methodology, since researcher should measure each concept by each method (Trochim, 2000).

Figure 4 involves three different traits (i.e., A, B, and C), each measured by three methods (i.e., 1, 2, and 3). The reliability diagonals suggest that the same construct is measured by same method, and that they share both method and construct variance. Hence, researcher would expect these correlations to be highest. Correlations corresponding to the same construct but different methods (i.e., validity diagonals) should be statistically significant and sufficiently large, which lends support to convergent validity (Campbell and Fiske, 1955). The validity diagonals should be the second highest. This would suggest that the construct as

a source of covariation among the measures is more important as compared to the method as a source of covariation. Hence, the measures are more influenced by what they are supposed to measure as compared to how they are measured.

Figure 4: An Illustration of the Multitrait-Multimethod Matrix Including Hypothetical Numbers

		Method 1			Method 2			Method 3		
	Traits	**A1**	**B1**	**C1**	**A2**	**B2**	**C2**	**A3**	**B3**	**C3**
	A1	(.89)								
Method 1	**B1**	.51	(.89)							
	C1	.38	.37	(.76)						
	A2	.57	.22	.09	(.93)					
Method 2	**B2**	.22	.57	.10	.68	(.94)				
	C2	.11	.11	.46	.59	.58	(.84)			
	A3	.56	.22	.11	.67	.42	.33	(.94)		
Method 3	**B3**	.23	.58	.12	.43	.66	.34	.67	(.92)	
	C3	.11	.11	.45	.34	.32	.58	.58	.60	(.85)

( ) : Relaibility diagonals
: Validity diagonals (i.e., monotrait-hetromethod values)
: Heterotrait-monomethod triangle
: Heterotrait-heteromethod triangle

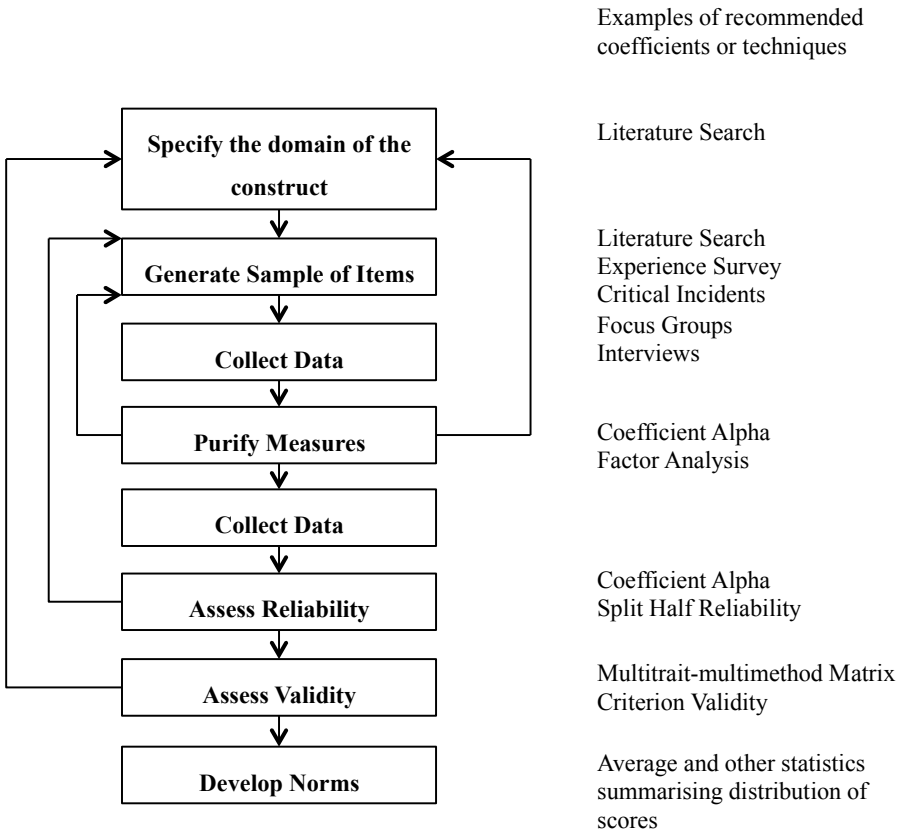
To support the discriminant validity, the validity correlations should be larger than the correlations of different constructs measured with the same method. Furthermore, the validity diagonals should also be higher than the values lying in the heterotrait-heteromethod triangles.

5.1.5 Marketing Constructs Validation Based On Classical Test Theory

In a seminal article, Churchill (1979) proposed a highly influential measurement paradigm in marketing. As seen below, his framework was substantially influenced by CTT. He (1979) outlined a complete framework combining the measurement process and measures of reliability as well as validity, which can be easily utilized by market researchers. The eight-

step procedures for developing construct measures and corresponding calculations are showed below (See Figure 5).

Figure 5: Suggested procedure for developing better measures (Churchill, 1979)



The first four steps aim to develop a reliable and valid measurement instrument or scale. It begins with specifying the domain of the construct, as the researcher must explicitly state what is included in the definition and what is excluded. The second step is to generate items which capture the domain as specified. The purpose for item generation in early stages (i.e., step 2) is to develop a set of items which tap each of the dimensions of the construct of interest. Then after constructing an item pool, researchers can edit these items or statements in order to make their wording as precise as possible.

After the item pool is carefully edited, further refinement would wait until actual data are collected. The average correlation of all the proposed items indicates the extent to which some common core is present. If all items in a scale are drawn from a single construct, responses to those items should be highly intercorrelated. Thus, Churchill (1979) emphasized

that coefficient alpha should be the first measure one calculates to assess the quality of the instrument. A low coefficient alpha indicates that the sample of items performs poorly in capturing the construct of interest, and that items that do not share the common core should be eliminated. In addition to coefficient alpha, researchers should also perform a factor analysis to determine the number of dimensions underlying the construct. After a second data collection involving the refined items, the coefficient alpha can also be used for the reliability estimation. In terms of the validity assessment, Churchill (1979) recommended the multitrait-multimethod matrix to estimate the construct validity (i.e., discriminant validity and convergent validity).

Gerbing and Anderson (1988) further developed the idea proposed by Churchill, and emphasized that reliability of each scale can be assessed only after unidimensional measurement has been acceptably achieved. The concept of unidimensionality implies that only the latent variable of interest (i.e. the underlying construct) is responsible for the observed covariation. Furthermore, the assessment of unidimensionality should employ confirmatory factor analysis, and item-total correlations and explanatory factor analysis (EFA) are better to be used as preliminary technique to generate preliminary scales or item pools rather than the assessments of unidimensionality (Gerbing and Anderson, 1988).

These two paradigms have been very influential in the marketing measurement literature and have made difficult measurement issues accessible to marketing researcher in general. However, there are several researchers that warn against or oppose some of the propositions put forward in these articles. For instance, Lee and Hooley (2005) addressed some problems associated with the use of techniques such as Cronbach's alpha and explanatory factor analysis (EFA) to identify relevant questions for measuring constructs. For example, a high coefficient alpha might imply not only a high level of reliability but also a worrying level of item redundancy (Lee and Hooley, 2005). Furthermore, the coefficient alpha is affected by the number of items in the scale to be tested (Lee and Hooley, 2005). Normally, a long scale would result in a high reliability index. Salzberger and Koller (2012) further suggested that using two unobservable components (true score and error score) provides little insight and defies empirical rejection.

5.1.6 Summary

Classical Test Theory has influenced perspectives of measurement not only in marketing and psychology but in most of social sciences. It has given rise to criteria to select “good” measures and to a number of beliefs about the way valid and reliable indicators or items should behave. For example, Nunnally (1978) warned that if correlations among measures are near zero, they measure different things. Furthermore, the reliability and validity indices regarding a scale or an item in Classical Test Theory are all defined as population level instead of individual level.

5.2 Generalizability Theory (GT)

5.2.1 Theory Overview and Definition

Generalizability theory (G-theory) offers an extensive conceptual framework and a powerful set of statistical procedures for addressing measurement issues, such as the dependability (i.e., reliability) evaluation of behavioral measurements (see Cronbach et al., 1972; Shavelson and Webb, 1991; Brennan, 2001). G-theory can be seen as a response to the limitation of the undifferentiated error component found in Classical Test Theory. Hence, it permits a multifaceted perspective on measurement error and its components, and enables an investigator to partition measurement error into multiple error sources in that Generalizability theory extends the decomposition of variance beyond the two components (true score and error) found in CTT and includes several sources of systematic error components (i.e., facets). Thus, these additional components of variation provide further understanding of the systematic sources of error and a prescription for how to control for these sources of error. A brief overview of Generalizability theory and its theoretical basis is provided in the following sections.

Facets and conditions

The sources of variation, in G-theory, are referred to as facets. It is crucial to determine which facet will serve as the object of measurement for the purpose of analysis. Rentz (1987) considered the object of measurement also as the facet of differentiation, since it refers to a set of objects to be compared in a study. In behavioral research, the person to whom a number/score is assigned to is typically considered to be the object of measurement (Shavelson et al., 1989). However, in marketing research, the object of measurement might not be always the person (Rentz, 1987), but might instead be a product or an ad campaign.

The remaining facets (e.g. rater, time, and setting) are considered to be sources of measurement error. Rentz (1987) refers these remaining facets as the facets of generalization. Facets of generalization contribute both systematic error and random error to observations in

a study. These are the facets over which researchers wish to generalize. On the other hand, within a specific research, a scale should minimize variance arising from these sources.

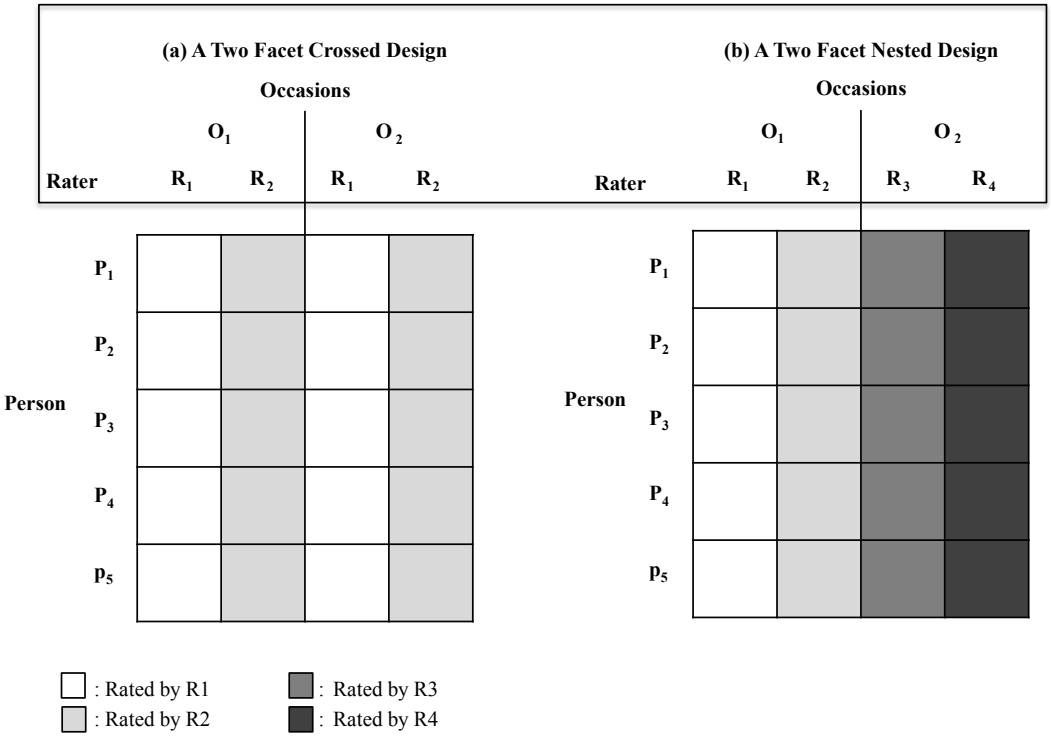
There can be several facets included in a G-study. A one-facet G-study includes only one source of measurement error (e.g. occasions, raters, or items). For instance, if the research includes three items rated by one observer in one situation at one point in time, this research design is considered as a one-facet design, because "item" is considered the only "facet" that allows generalization. A facet is analogous to a factor in analysis of variance; and conditions, or levels, of a facet are analogous to levels of a factor in analysis of variance (Rentz, 1987). The conditions of a facet may be either fixed or random. However it might be difficult to determine whether items should be considered random or fixed. Items usually are considered randomly chosen from a universe of items, even though they in most cases obviously are not. Few researchers would suggest that the items in a particular instrument exhaust the set of possible items, suggesting that the facet items should be treated as fixed. However, in practice, items usually are considered random facets. It is not difficult to assess the consequence of treating facets as fixed or random in the analysis of a generalizability study[18].

Facets may also be crossed or nested. Two facets are said to be crossed if every measurement condition of the first facet occurs in combination with every measurement condition of the second facet. A facet is said to be nested within a second facet if different sets of measurement conditions of the first facet occur in combination with each measurement condition of the second facet (Crocker and Algina, 1986). For example, if we have a two facets[19] $p * r * o$ (i.e., person by rater by occasion) measurement design, five persons (i.e., the object of measurement) are rated by two raters in two occasions (see Figure 6).

[18] See Rozeboom (1978) for a criticism of the random sampling assumption found in generalizability theory.

[19] The object of measurement, persons, is not referred to as facets. Hence, what appears to be three factors would be referred to as one object of measurement (persons) and two facets of generalization (raters and occasions).

Figure 6: Outline of a Two-Facet (a) Crossed, and (b) Nested Design



The left table (a) is considered as crossed design, since the two facets (i.e., raters and occasions) are crossed with each other, (e.g., O_1R_1, O_1R_2, O_2R_1, O_2R_2). In contrast, the right table (b) is considered as nested design, because raters 1 and 2 code behavior on occasion 1 (e.g., O_1R_1, O_1R_2) and two different raters 3 and 4 code behavior on occasion 2 (e.g., O_2R_3, O_2R_4). In this case, the facet rater, r, is nested within facet occasion, o, which can be denoted as $r{:}\,o$ or $r(o)$. In the two-facet design, G-theory is able to evaluate the differences not only among raters but also among occasions.

5.2.2 Generalizability Study and Decision Study

Generalizability theory recognizes that an assessment might be adapted for particular decisions and so distinguishes a Generalizability (G) study from a Decision (D) study. In order to evaluate the dependability (i.e., reliability) of behavioral measurement, a G-study is designed to isolate particular sources of measurement error. A G-study is primarily concerned with the extent to which a sample of measurements generalizes to a universe of measurements. Thus, if the facets of measurement error are reasonably and economically feasible, then they should be isolated and estimated in the G-study (Webb et al., 2006)

Subsequently, a D-study uses the information provided by the G-study to design the best possible application of the measurement for a particular purpose. In the D-study, the researcher or decision maker might want to make two types of decisions based on a behavioral measurement: relative or absolute (Shavelson et al., 1989). A relative decision focuses on the rank order of persons, while an absolute decision focuses on the level of performance. For instance, five advertisements might be ranked according to their scores on a copy test and the two highest ranked advertisements considered for further development. This decision is a relative one. If an advertisement will be considered for further development only if the score on a copy test exceeds some minimum score, then the decision is an absolute one.

The purpose of a G-study is to help the researcher to plan a D-study that will have adequate generalizability. To minimize error and maximize reliability, the decision maker would use the information/generalizable data provided by a G-study to evaluate the effectiveness of alternative. Thus the design of the G-study needs to anticipate the full variety of designs that may be used for the D-study. However the D-study may contain fewer (or more) conditions, depending on the purpose of the study and the results of the G-study (Rentz, 1987). For example, if the G-study shows that certain facets contribute little to the overall error, the number of conditions of those facets can be reduced in subsequent D-studies with little loss of generalizability. Resources might be better spent on increasing the sample of conditions that contribute larger amounts of variance to the overall error so that generalizability is increased. The ability to predict and control the sources and magnitude of measurement error in subsequent studies is unique to G-theory and should be of great practical importance to marketing researchers.

It is not possible to classify a study as a G-study or a D-study based on its design alone; the purpose of the investigator is the determining factor (Crocker and Algina, 1986). Suppose that a researcher tests 200 employees randomly selected from state-owned enterprises and 200 employees randomly selected from private enterprises, using a single standardized job satisfaction test. If the purpose is to determine whether the test is equally reliable for employees in both types of enterprises, this would be classified as a G-study. Conversely, if the investigator wants to compare the mean job satisfaction levels of the two groups and to draw conclusions about possible differences of the two working conditions, this would be

considered a D-study. In sum, G-study is associated with the development of a measurement procedure whereas D-study then applies the procedure.

5.2.3 G-study: Universe of admissible observations and Universe of generalization

In G-theory, a behavioral measurement observation (e.g., a test score) is conceived of as a sample from a universe of admissible observations. This universe contains all possible observations of the object of measurement that a decision maker considers to be acceptable substitutes for the observation in hand (Shavelson and Webb, 1991). Shavelson and Webb (1991) further defined a universe of admissible observations as all possible combinations of the conditions of the facets (Shavelson et al., 1989).

The universe of generalization is a fundamental notion in G-Theory. It refers to the set of facets and their conditions to which a decision maker wants to generalize (Shavelson et al., 1989). Normally, marketing researchers are rarely interested in generalizing a measurement over only the particular occasion on which the measurement is taken, but would rather be interested in generalizing to the set of all such occasions (usually within some time interval). Hence, there is a universe of occasions to which researchers wish to generalize. Similarly, researchers might wish to generalize over a universe of items, interviewers, situations of observation, etc. The set of all such conditions of measurement over which the investigator wishes to generalize is the universe of generalization (Rentz, 1987). The universe of generalization may differ among studies according to the purposes of the studies. Therefore a researcher must define the universe unambiguously by specifying precisely the conditions of measurement over which he or she intends to generalize in a particular study.

There is a universe of observations that would have provided a usable basis for the decision. Thus the ideal datum on which to base the decision would be something like the person's mean score over all acceptable observations, which is called the person's "universe score"[20] (Cronbach et al., 1972). The universe score, denoted as μ_p, refers to the underlying construct and can be defined as the expected value of his or her observed scores over all observations in the universe of generalization (Shavelson et al., 1989).

[20] The universe score refers to the latent variable of interest in G theory.

Components of observed score

After the decision maker specifies the universe of generalization, an observed measurement can be decomposed into a component for the universe score and one (or more) error components (Shavelson et al., 1989). Classical Test Theory and Generalizability Theory both can be used in a one-facet design, $p * i$ (person by item), for reliability estimation. For example, a traditional internal consistency analysis contains the object of measurement, the person, and one facet, the items. However, if we design a two-facet crossed measurement study, $p * i * o$ (person by item by occasion), where items and occasions have been randomly selected[21], an observed score, X_{pio}, can be decomposed into the following sources of variance (see Table 3):

Table 3: Components of the observed score in a two-facet crossed design

$\boldsymbol{X_{pio} = \mu}$	**Grand mean**
$\boldsymbol{+\mu_P - \mu}$	Personal effect
$\boldsymbol{+\mu_i - \mu}$	Item effect
$\boldsymbol{+\mu_o - \mu}$	Occasion effect
$\boldsymbol{+\mu_{Po} - \mu_P - \mu_o + \mu}$	Person * Occasion effect
$\boldsymbol{+\mu_{Pi} - \mu_P - \mu_i + \mu}$	Person * Item effect
$\boldsymbol{+\mu_{oi} - \mu_o - \mu_i + \mu}$	Occasion * Item effect
$\boldsymbol{+\ residual}$	$p * o * i$, error

P = person; I = items; O = occasions; μ = constant

Except for the grand mean, the distribution of each component has a mean of zero and a variance component σ^2. The variance component for the person effect, σ_p^2, is called the

[21] This design is assumed to generalize over all admissible test items and occasions taken from an indefinitely large universe.

universe score variance, and the variance component for the other effects are considered error variation. In this two-facet crossed design, G-theory addresses the interaction variances (e.g., between person and item, person and occasion, occasion and item) and enables researcher to interpret these variances respectively. For example, the person*occasion effect shows the inconsistency in relative standing of the person from one occasion to the next. Similarly, the person*item variance indicates the inconsistency in relative standing of the person from one item to another. The item*occasion effect can be interpreted as the inconsistency in items' average effects on persons from one occasion to another. However, it would be difficult to interpret these interaction effects in a higher order facet design, such as a five-facet design (Peter, 1979). The residual variance component, $\sigma^2_{pio,e}$, reflects the person*item*occasion interaction confounded with residual error. The variance of the observed score, σ^2_{pio}, is showed as below:

$$\sigma^2_{pio} = \sigma^2_p + \sigma^2_i + \sigma^2_0 + \sigma^2_{pi} + \sigma^2_{po} + \sigma^2_{oi} + \sigma^2_{pio}$$

Each variance component can be estimated by the analysis of variance, ANOVA. The relative magnitudes of the estimated variance components, except for the universe score variance, provide information about potential sources of error influencing a behavioral measurement.

5.2.4 D-study: Generalizability coefficient and Dependability index

Generalizability theory assumes that the universe score is a mean (or sum) score over samples of conditions of the measurement. Researchers are rarely interested in a person's response to an individual item, but in the person's mean (or sum) score over samples of all items in the universe of generalization (Cronbach et al., 1972). Hence, in a D-study, decisions usually will be based on the mean over multiple observations rather than on a single observation. To emphasize this difference, we can use X_{pIO} to represent the mean score over a sample of items and occasions, instead of X_{pio} reflecting the mean score over a single item and occasion.

Furthermore, in a G-study Generalizability theory provides researchers the interpretation of variance components and measurement error, whereas in the D-study it provides summary

coefficients that are analogous to the reliability coefficient in CTT (Shavelson and Webb, 1991). However, the focus of G-theory is still on analyzing components of variance, rather than on these summary coefficients. As mentioned above, there are two types of decision in a D-study, relative and absolute decision. The error variance is defined differently for each kind of decision (Shavelson and Webb, 1991). Similar as with the error variances, the summary coefficients also differ in relative and absolute decisions. G-theory distinguishes between a Generalizability Coefficient for relative decisions and an Index of Dependability for absolute decisions.

For relative decisions, the relative error variance σ_δ^2, in a two-facet $p * i * o$ crossed design with random effects, is defined as below:

$$\sigma_\delta^2 = \sigma_{pI}^2 + \sigma_{pO}^2 + \sigma_{pIO,e}^2 = \frac{\sigma_{pi}^2}{n_i'} + \frac{\sigma_{po}^2}{n_o'} + \frac{\sigma_{pio,e}^2}{n_i' n_o'}$$

The main effects of item and occasion do not enter into the error for relative decisions. All people respond on both occasions and items, so any difference in occasions or items affects all persons and doesn't change rank order (Shavelson and Webb, 1991). The corresponding Generalizability Coefficient, E_ρ^2, is the ratio of universe score variance to the expected observed score variance (i.e. an intraclass correlation). For the relative decision in a $p * i * o$ random-effects design, the Generalizability coefficient can be expressed as below.

$$E_\rho^2 = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_\delta^2}$$

In contrast, an absolute decision focuses on the level of an individual's performance independent of others' level. For the absolute decision in a $p * i * o$ random-effects design, the variance of the absolute error σ_Δ^2 can be expressed as follows:

$$\sigma_\Delta^2 = \sigma_I^2 + \sigma_O^2 + \sigma_{pI}^2 + \sigma_{pO}^2 + \sigma_{IO}^2 + \sigma_{pIO,e}^2$$

$$= \frac{\sigma_i^2}{n_i'} + \frac{\sigma_o^2}{n_o'} + \frac{\sigma_{pi}^2}{n_i'} + \frac{\sigma_{po}^2}{n_o'} + \frac{\sigma_{io}^2}{n_i' n_o'} + \frac{\sigma_{pio,e}^2}{n_i' n_o'}$$

The main effect of items (e.g. how difficult an item is) and occasions does affect the object of measurement, even though neither changes the rank order. The Dependability Index in this $p * i * o$ random-effects design is:

$$\phi = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_\Delta^2}$$

The difference between σ_δ^2 and σ_Δ^2 is that σ_δ^2 does not include sources of variance common to each person (e.g., the main effect of item and occasion as well as their interaction effect) whereas σ_Δ^2 does. Accordingly the Dependability index would be lower than the Generalizability coefficient. Most marketing decisions are relative decision, so the relative error variance is appropriate in most cases (Rentz, 1987).

5.2.5 Reliability

Cronbach et al. (1972) broadened the conception of measurement reliability by introducing the notion of generalizability, which is at the heart of G Theory. Briefly, measurement quality is evaluated in terms of the ability to make inferences from (a) scores based on a limited number of observations to (b) scores based on nearly unlimited number of observations: "The question of reliability thus resolves into a question of accuracy of generalization, or generalizability" (Cronbach et al., 1972, p15).

Furr and Bacharach (2008) pointed out that the reliability in Generalizability theory follows the domain sampling theory. Domain sampling theory of reliability, developed in the 1950s, is an alternative to Classical Test Theory. Both domain sampling theory and CTT will provide similar results in terms of reliability estimation, but they differ in terms of the assumptions needed to provide the estimate (Ghiselli et al, 1981). CTT reliability rests on the assumption that it would be possible to create two tests that are parallel to each other. Domain sampling

theory, however, relies on another assumption that items on any particular test represent a sample from a large indefinite number or domain of potential test items.

Furthermore, G-theory enables researchers to achieve a higher dependability (i.e., reliability) through redesign. Like Spearman-Brown "prophecy formula" which is used to predict reliability as a function of test length in CTT, decision makers in G theory can determine that how many occasions, test forms, and administrators would be needed to obtain dependable scores. Moreover, G theory also allows the decision maker to investigate the dependability of scores for different kinds of interpretations. In addition to the interpretation about relative standing of respondents, which is the only issue addressed in CTT, G-theory provides the information regarding the absolute level interpretation.

5.2.6 Validity

Generalizability theory does not easily fit in the traditional distinction between reliability and validity. Cronbach et al. (1963) addressed the validity issue as follows:

> "Our rationale requires the investigator to start his G study by defining the universe that interests him, and then to observe under two or more independently selected conditions within that universe. The calculations tell him how well the observed scores represent the universe scores. Since the universe is a construct that he introduces because he thinks it has explanatory or predictive power, an investigation of generalizability is seen to be an investigation of the 'construct validity' of the measure. The theory of 'reliability' and the theory of 'validity' coalesce; the analysis of generalizability indicates how validly one can interpret a measure as representative of a certain set of possible measures." (Cronbach et al. 1963, p157)

The relationship between generalizability theory and validity has been addressed extensively by Kane (1982). Since the universe score for each object of measurement equals the value of the object, a measurement procedure is valid to the extent that it estimates the universe scores accurately. For a measurement procedure consisting of random sampling from the universe of generalization, the observed score is an unbiased estimate of the universe score. Kane (1982) pointed out that because the generalizability coefficient indicates how accurately universe

scores can be inferred from observed scores, it can be interpreted as a validity coefficient. Therefore, if a latent construct is clearly specified in terms of a universe of generalization and if random samples could be drawn from this universe, validation would be relatively straightforward.

However, the universe of generalization is usually not clearly defined (Kane, 1982). Hence, this complicates the analysis of validity. Although a generalizability coefficient can be an index of validity, most estimated generalizability coefficients would not be qualified as validity coefficients (Kane, 1982). The interpretation of the generalizability coefficient as a validity coefficient depends on the extent to which the universe of generalization is defined adequately and the extent to which the sample from the universe is random. An adequate definition of the universe would contain a complete delineation of the facets and conditions of facets in the universe. Defining the conditions (or levels) of each facet is similar to defining the domain of items in Classical Test Theory. In a particular G study it is likely that only a subset of these facets would be investigated. In this case, it would be better to just interpret Generalizability coefficient as a reliability coefficient, indicating the extent to which one can generalize over the particular subset of facets.

5.2.7 Summary

Generalizability theory, a rival candidate to replace Classical Test Theory in measurement (Conger 1981), is a comprehensive and flexible method of assessing and improving the dependability of measurements. Although the higher order interactions are difficult to interpret and might limit the usage of GT in marketing (Peter, 1979), these interactions should not be ignored as they are in traditional methods of reliability assessment. The power of ANOVA enables researchers to decompose the variance and hence provide a better analysis of various sources of measurement error. Furthermore, as Rentz (1987) suggested, “measurement occasion” contributes significantly to measurement error in marketing applications (also see Peter, 1979) in addition to “the items” in a scale. Hence, the Generalizability Theory should be promoted in marketing measurement to ensure that several sources of measurement error can be investigated (e.g., the joint impact of error from items and occasions).

5.3 Item Response Theory (IRT)

5.3.1 Theory Overview and Definition

Prior to the introduction of Item Response Theory (IRT), the statistical measurement of empirical data was based entirely on CTT. Lord and Novick's textbook (1968) provided the first comprehensive description of Item Response Theory and its models, which are based on the cumulative normal distribution (normal ogive). Later Allen Birnbaum (1968) substituted the normal ogive model with the logistic model due to its computationally simple form, and extended the idea suggested by Lord (1968). Another key figure of IRT development is Rasch (1960), who developed a different, but influential IRT model (i.e., the Rasch model).

Following the pioneering work of Birnbaum, Rasch and others, considerable research efforts have been devoted to IRT. In particular, the increased availability of computational capacity offered by the rapid development of computers overcomes the initial disadvantage of IRT in terms of demanding computational requirements. All these efforts have made research on IRT one of the most active areas in the social science. Currently, IRT has gradually become a much used theoretical basis for measurement (e.g., Baker and Kim, 2004; Jong et al., 2008; De Ayala, 2009).

Item Response Theory (IRT), also referred to as latent trait models, encompasses a diverse set of models designed to represent the relation between an individual's item response and an underlying latent variable θ (often called "ability" or "trait") (Van der Linden and Hambleton, 1997). It assumes that one (or more) latent variables are responsible for the responses to test items (i.e., measures). Hence, variation among individuals on the latent variable explains the observed covariation among item responses (Green, 1954). In IRT models, the relation between the position of individuals on the latent variable(s) (i.e., item location parameter) and the item responses is depicted by a statistical model that describes the probability of an item response as a function of the latent variable(s). Hence, IRT specifies that the performance of an examinee on a test item can be predicted (or explained) by a latent variable or a set of latent variables. Furthermore, the relationship between an examinee's item performances and the latent variable underlying item performances can be described by a function called an item characteristic function or item characteristic curve

(ICC). Although early IRT models emphasized dichotomous item formats and usually assumed unidimensional models, extensions to other item formats and multidimensional models have enabled IRT applications in many areas.

A particular aspect of IRT models is the reliance on the property of invariance, which makes IRT different from CTT (Hambleton et al., 1991). The invariance property implies that the trait level θ and item parameters are independent to each other. Thus, participants can be compared independently of the items involved. IRT achieves this by (1) incorporating information about the items into the trait-level estimation process and (2) incorporating information about the examinees' trait levels into the item-parameter-estimation process. However, this property only holds when the model perfectly fits the data (Hambleton et al., 1991). Since researchers cannot confidently state that the model fits data, the only thing they can do is to estimate the model-data fit and infer the corresponding "degree" to which invariance holds (Hambleton et al., 1991).

Basic Item Response Theory Assumptions

The mathematical model employed in IRT specifies that an examinee's probability of answering or endorsing a given item correctly depends on the examinee's ability or abilities and the characteristics of the item. IRT models include a set of assumptions about the data to which the model is applied. A common assumption of IRT models is the one of "appropriate" dimensionality (Embretson and Reise, 2000). In the overwhelming majority of applications, IRT models assume unidimensionality, that is, only one trait is measured by a set of items in a test (Hambleton et al., 1991). Thus the probability of endorsing an item solely depends on the latent trait of interest, and no other traits are proposed to account for this probability. Models fulfilling this assumption are referred to as unidimensional models. However, there are also multidimensional models, introducing more than one latent trait to explain the test performance. Although these multidimensional IRT models do receive increased attention in the literature (e.g., Ackerman, 1994; Reckase, 1997), IRT models for the most part still assume unidimensionality.

Another important assumption is the one of local independence of items. Local independence states that, when the latent trait influencing test performance is held constant, participants'

responses to any pair of items are statistically independent[22] (Hambleton et al., 1991). Thus, after taking the trait into account, no relationship exists between participants' responses to different items, implying that the latent trait specified in the model is the only factor influencing participants' responses to test items. Based on this assumption, Item Response Theory formally defines what the latent variable is.

The unidimensionality assumptions and local independence assumption would be the same given only one latent trait. However, in the multidimensional models, where the unidimensionality assumption turns irrelevant, local independence will still be useful to determine whether several traits are responsible for the covariance under item performances. The reason is that local independence is relevant regardless how many latent traits are specified to influence the item scores and hence local independence does not rely on unidimensionality (Hambleton et al., 1991). When all the latent traits influencing performance have been held constant, local independence can still be confirmed if participants' item scores are statistically independent.

Another important assumption made in all IRT models is that the item characteristic function reflects the true relationship among the latent variables (traits) and the observable variables (item responses). Furthermore, IRT also requires assumptions about the item characteristics that are relevant to an examinee's performance on an item. The major distinction among the IRT models is found in the number and type of item characteristics assumed to affect examinee's performance.

5.3.2 Different IRT Models and Item Parameters

As previously mentioned, there are many different types of IRT models. For example, IRT models might differ as pointed out above depending on the number of dimensions they account for. Some models are unidimensional whereas other models are multidimensional. Furthermore, IRT models can differ in how many scored responses they account for. For instance, the typical multiple choice item is dichotomous and responses are scored as either correct or incorrect (e.g., agree or disagree; yes or no). Thus, there are only two response

[22] The probability of solving any item is independent of the outcome of any other items, controlling for item parameters and the trait level.

categories for a dichotomous item. Within dichotomous item response models, a primary distinction is the number of item parameters used. The most popular IRT models are the one, two, and three-parameter(s) logistic models. The one-parameter model also refers to the Rasch model or simple logistic model.

Another class of models apply to polytomous item responses (i.e., three or more response categories), where each response has a different score value. For instance, a Likert-type item may have five different response categories. In these polytomous models, each individual response category (e.g., from 1 to 5) is considered explicitly, thus accounting for the discrete character of the responses (see van der Linden and Hambleton, 1997; Roberts et al., 2000). Unlike the dichotomous model where only one item characteristic curve is considered, the polytomous model considers five item characteristic curves for a five-category polytomous item (see Figure 12 in Appendix). For the polytomous items, there also exist different types of models, such as graded model (ordered responses, like Likert scale), nominal model (no pre-specified order).

Polytomous models are more complicated than dichotomous models, and since the rationale for Item Response Theory can be presented more easily for the dichotomous models, I will focus only on dichotomous models in my presentation of Item Response Theory.

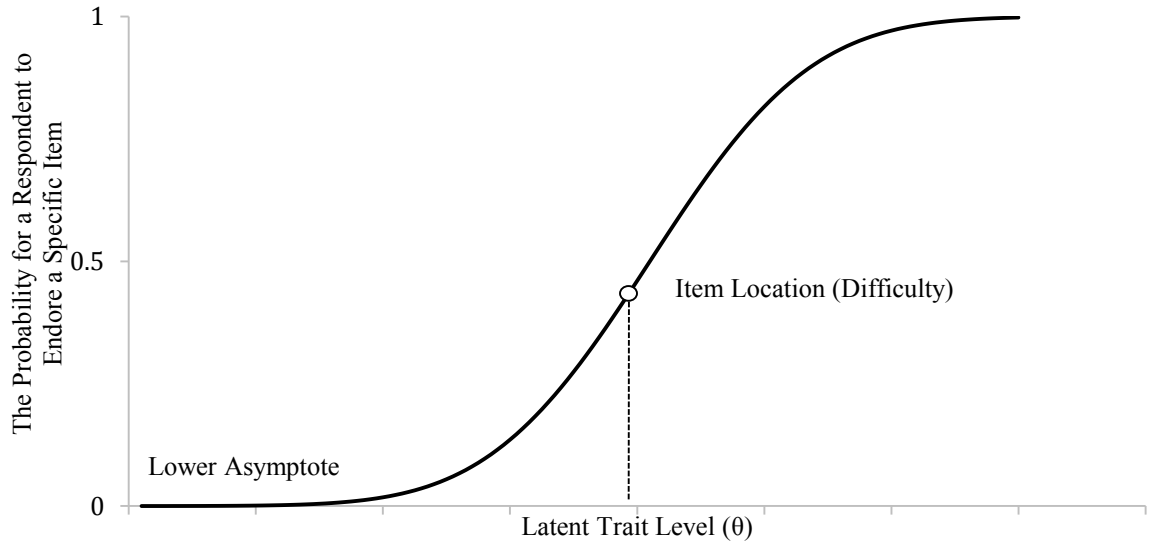
One-parameter logistic model

The one-parameter logistic model is one of the most widely used dichotomous item response models. The item characteristic curve for a one-parameter logistic model is given in Figure 7.

The item characteristic curve (ICC) is a central concept in IRT. An ICC plots the probability of endorsing an item as a function of the latent trait being measured (denoted by θ). For dichotomous items, the ICC regresses the probability of answering an item on the trait level; and for polytomous items, the ICC regresses the probability of item responses in each category[23] on the trait level.

[23] It is the probability of choosing one specific response category. For example, given a trait level, an examinee may have 1% possibility to choose "Strongly Disagree", 3% possibility to choose "Disagree", 10% possibility to endorse "Neutral", but 50% possibility to endorse "Agree".

Figure 7: Item Characteristic Curve for One-Parameter Logistic Model



According to Lord (1980) there are two acceptable interpretations. The first interpretation implies that the probability of responding correctly is interpreted as the probability that a random chosen member from a homogeneous subpopulation will respond correctly to an item. Members in the homogeneous subpopulation are the same in terms of the latent trait (i.e. have the same latent trait score). The second acceptable interpretation refers to a subpopulation of items all of which have the same ICC. The probability of responding correctly is then interpreted as the probability that a specific examinee will endorse an item randomly chosen from the subpopulation of items.

In most applications of Item Response Theory, ICC is assumed to have an S-shape. Figure 7 shows that as the score on the latent trait increases, so does the probability of endorsing a specific item. The importance of the ICC is that it permits researchers to see how the probability of endorsing an item depends on the latent variable. The one-parameter logistic model for a given item i can be expressed as:

$$P_i(\theta) = \frac{e^{(\theta - b_i)}}{1 + e^{(\theta - b_i)}} \qquad i = 1, 2, \ldots, n$$

In the above equation, the observed variables do not figure directly. Instead the probability of endorsing an item does appear in the equation. Furthermore, $P_i(\theta)$, represented by the S-shape curve with values between 0 and 1 over the latent trait scale, is the probability that a

randomly chosen participant with the trait level θ endorsing a specific item i or answering the item i correctly. n in the equation represents the number of items in the test and e is a transcendental number whose value is approximately 2.718.

The difficulty or location parameter, b_i, is the only item parameter in this model. When an item is dichotomously scored, the item difficulty refers to the proportion of examinees who answer the item correctly or who endorses the item (Crocker and Algina, 1986). Statistically, the b_i parameter for an item i represents the point on the latent trait scale where the probability of a correct response is 0.5 (See Figure 7). This parameter is a location parameter, indicating the position of the ICC in relation to the trait scale. A higher value of the b_i parameter implies that a higher level of the trait (or ability) is required for an examinee to have a 50% chance of getting the item right. Thus, b_i does reflect whether the item is difficult or simple.

In the one-parameter models, the ICCs differs only in their location parameters (Hambleton et al., 1991), and thus item difficulty is the only item characteristic that influences participant's performance. All items are equally discriminated (i.e., have a fixed slope for all items). Furthermore, the lower asymptote of this ICC is zero, which states that examinees with very low ability have almost zero probability of endorsing the item. No allowance is made for the possibility that low-ability examinees may guess, as they are likely to do on multiple-choice items.

The one-parameter model is based on restrictive assumptions (e.g., all items are equally discriminated; lower asymptote is zero). The appropriateness of these assumptions depends on the nature of the data and the intended application. The one-parameter logistic model is often called the Rasch model to honor its developer. While the form of Rasch's model[24] is different from that presented here, the one-parameter logistic model is mathematically equivalent to Rasch's model (Hambleton et al., 1991). However, as the Rasch model requires item discrimination to be equal across items, the proponents of the Rasch model prefer to

[24] In the original Rasch model, a person is characterized by a level on a latent trait ξ, and an item is characterized by a degree of difficulty δ. The probability of an item endorsement is a function of the ratio of a person's level on the trait to the item difficulty ξ/δ. For a particular item, a trace line (probability) function increases from zero to one with trait level: $T = \frac{\xi}{\xi+\delta}$.

view it as a completely different approach from general IRT models (See e.g., Andrich, 2004; Salzberger and Koller, 2011). Compared to the general IRT model which seeks an optimal description of the data, the Rasch model takes precedence over the data. If the data do not fit the Rasch model, most researchers would resort to a more general model, whereas proponents of the Rasch model would reject the measurement.

Two-parameter logistic model

Item characteristic curves for the two-parameter logistic model developed by Birnbaum (1968) are given by the following equation:

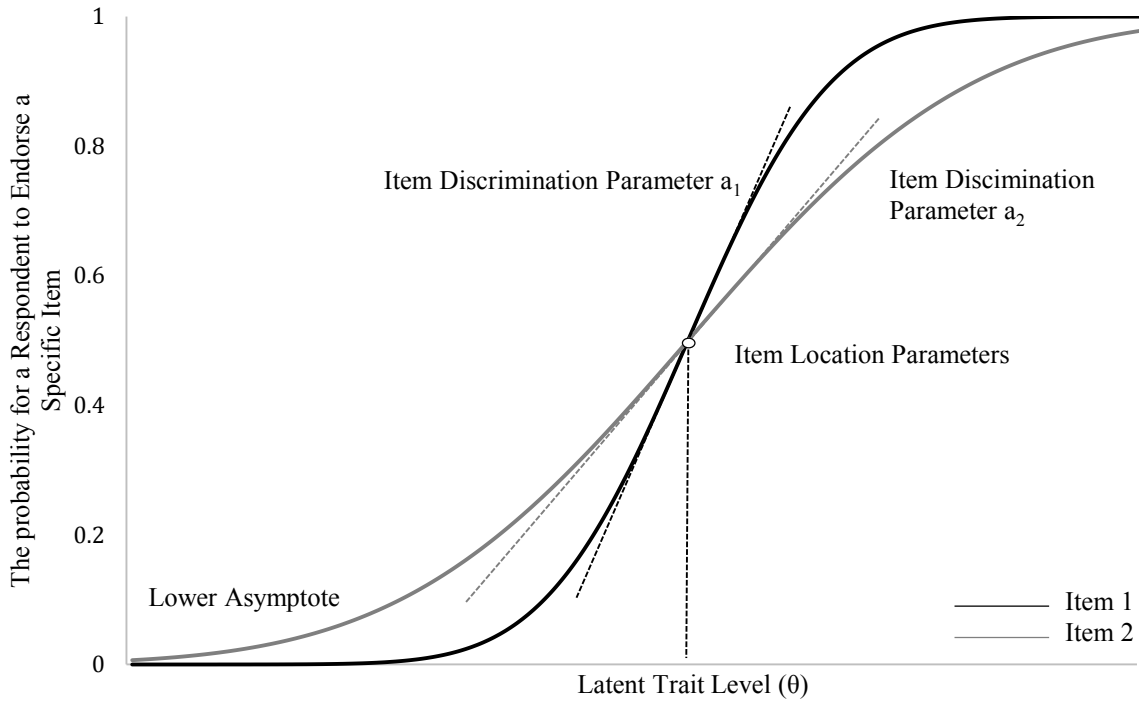
$$P_i(\theta) = \frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}} \qquad i = 1, 2, \ldots, n$$

The two-parameter logistic model resembles the one-parameter model except for the presence of two additional elements. The factor D is a scaling factor introduced to make the logistic function as close as possible to the normal ogive function (Hambleton et al., 1991). It has been shown that when D = 1.7, values of $P_i(\theta)$ for the two-parameter normal ogive and the two-parameter logistic models differ in absolute value by less than 0.01 for all values of θ (Hambleton et al., 1991).

The second additional element of the two-parameter model is the parameter a_i, the item discrimination parameter. This parameter represents an item's ability to differentiate between people with continuous trait levels, and shows how rapidly the probabilities change with trait level. The a_i parameter is the slope of the ICC at the point b_i on the trait scale. In Figure 8, two item characteristic curves share the same location parameters[25], but differ in the discrimination parameters. Items with steeper slopes (a_1) are more useful for separating examinees into different trait levels as compared to items with less steep slopes (a_2). Furthermore, an item's ability to discriminate between people with similar trait levels is highest in the θ region, i.e. near location parameter (Fraley et al., 2000).

[25] In the two-parameter models, location parameters and discrimination parameters all can be different from one item to another item.

Figure 8: Item Characteristic Curves for Two-Parameter Logistic Model



In the two-parameter model, item difficulty and item discrimination parameters can determine the shape of the item characteristic curve. Researchers would be better to choose the items that have a relatively high discrimination value, since they are the better indicators of a latent trait. However, the two-parameter model also makes no allowance for guessing behavior. The assumption of no guessing is not plausible to be met in multiple-choice items, especially when a test is not too difficult for the participants.

Three-parameter logistic model

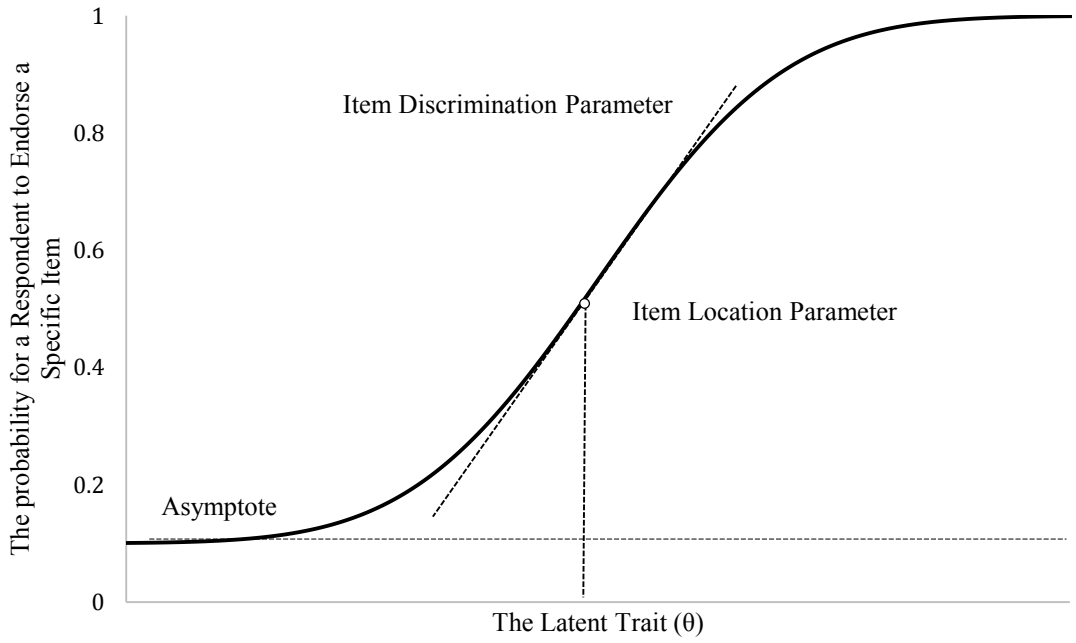
The three-parameter logistic model adds one more parameter c_i, pseudo-chance-level parameter (see Figure 9), and can be represented as below (Hambleton et al., 1991):

$$P_i(\theta) = c_i + (1 - c_i)\frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}} \qquad i = 1, 2, \ldots, n$$

This parameter provides a possibly nonzero lower asymptote for the item characteristic curve and represents the probability of examinees with low ability answering the item correctly. The parameter c_i is incorporated into the model to explain performance at the low end of the

trait continuum, where guessing is a factor in test performance. Typically, c_i assumes values that are smaller than the value that would result if examinees guessed randomly on the item (Hambleton et al., 1991). As Lord (1974) has noted, this phenomenon probably can be attributed to the ingenuity of item workers in developing attractive but incorrect choices. For this reason, c_i should not be called the "guessing parameter" (Hambleton et al., 1991).

Figure 9: Item Characteristic Curve for Three-Parameter Logistic Model



In addition to the Rasch (one), two, and three-parameter(s) logistic models, many other IRT models exist (see e.g., Masters and Wright, 1984; McDonald, 1989; Spray et al., 1990), differing in the mathematical form of the item characteristic function and/or the number of parameters specified in the model. Choosing a specific model is crucial to researchers. Multiple concerns regarding how to choose an appropriate model have been put forward including (1) whether items need to be equally discriminated, (2) the scale property of observed variables (e.g., either nominal or ordinal, dichotomous or polytomous), (3) the purpose of the study at hand, and (4) the data demands of individual IRT models (Embretson and Reise, 2000). Furthermore, as Hambleton et al. (1991) suggested, researchers should verify the model by examining how well it "explains" or fits the observed test results after choosing a specific model.

5.3.3 Item Parameter and Latent Trait Estimation

The purpose of giving a test is to determine a person's trait level from his or her responses to the items. In Item Response Theory, estimating θ is to determine what level of θ is most likely to explain the participant's responses, assuming the properties of the items and knowledge of how item properties influence behavior (i.e., item parameters) are known (Hambleton et al., 1991). Hence, estimating item parameters in IRT is based on the assumption that we know the participant's trait level θ. However, in reality we only have one known quantity, responses from the participant, but two unknown quantities, item parameters and the trait level for the participant. Under this situation, thanks to the development of computational capability, item parameters and trait level can be estimated simultaneously with numerical approaches.

In order to produce the "best fitting" curve, maximum likelihood could be used to estimate the item parameters and trait level (Hambleton et al., 1991). Estimates of parameters are determined by maximizing the likelihood of realizing sample residuals and this is obtained through computational iteration. Thus, finding the IRT trait level and item parameters require a search process (Embretson and Reise, 2000). Several computer programs (e.g., BICAL, LOGIST, BILOG, RUMM, MULTILOG, MIRTE, PARSCALE etc.) are available for estimation. Different programs are more suitable for different IRT models and hence choice of software will depend on the specific IRT model one wants to employ.

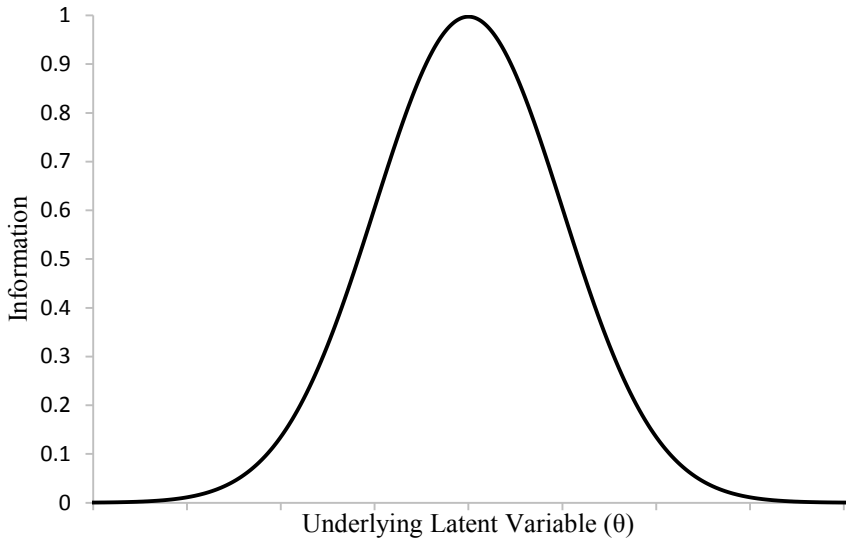
5.3.4 Reliability

IRT extends the concept of reliability from a single index to a function called the information function. The information function, which can be calculated[26], is typically presented in a graph (see Figure 10). The information function indicates that the range of trait level θ over which an item (or a scale) is most useful for distinguishing among participants. It characterizes the precision of measurement for persons at different levels of the underlying construct, with higher levels of discrimination providing more precision (Reeve, 2002).

[26] $I_i(\theta) = \frac{[P_i'(\theta)]^2}{P_i(\theta)Q_i(\theta)}$, where $I_i(\theta)$ is the "information" provided by item i at θ, $P_i'(\theta)$ is the derivative of $P_i(\theta)$ with respect to θ, $P_i(\theta)$ is the item response function, and $Q_i(\theta) = 1 - P_i(\theta)$.

The item's difficulty parameter determines where an item information function is located (Flannery, Reise, and Widaman, 1995). The higher the item's discrimination, the more peaked the information function will be. The information function can also be connected with the standard error of measurement towards the θ estimates. At each level of the underlying trait θ, the information function is approximately equal to the expected value of the inverse of the squared standard errors of the θ estimates (Lord, 1980). The smaller the standard error of measurement, the scale offers more information or precision about θ.

Figure 10: Item Information Curve for an Item



Although Figure 10 is just the information function for one item, with the assumption of local independence, different item information values can be summed across all of the items in a scale to form a test information curve[27] (Lord, 1980). Tests including more items measuring the latent trait provide more precise estimates as compared to a a single item. The more items are included in a test, the greater amount of information the test can provide (Baker, 2001). However, IRT offers also an alternative strategy to improve precision. Instead of increasing the length of a test, researchers can also increase the discriminative ability of each item to gain more information.

[27] The test information curve is defined as: $I(\theta) = \sum_{i=1}^{N} I_i(\theta)$. N is the number of items in the test.

5.3.5 Validity

The invariance property of IRT will only be satisfied if the fit between the model and test data is satisfactory, thus in IRT applications researchers are required to assess the fit of the model to the data. However, even though the topic of model-fit is an active area of current research, there are no unified procedures assessing model-fit for IRT models (Embretson and Reise, 2000). Hence, judging fit in IRT models calls for a variety of procedures, and ultimately, scientists have to use their own judgment.

Hambleton and Swaminathan (1985) recommended that judgments about the validity of a model to the test data should be based on: (1) the validity of the relevant assumptions, (2) the extent to which the expected properties of the model (e.g., invariance of item parameters and trait level θ) are obtained, and (3) the accuracy of model predictions using real and predicted test data (i.e., model-data fit). However, they also admitted that the extent to which the expected properties (e.g., invariance) holds depends on the model-data fit. Thus researchers normally employ only the first and the third criteria to examine their IRT models. Later, other researchers (see e.g., Embretson and Reise, 2000) also pointed out that the relevant validity assumptions and the goodness-of-fit of IRT models are more essential as compared to the expected properties outlined above and hence the first and third criteria need to be the primary goal of the test. Hence I will also focus the attention to the first and third criteria in the subsequent presentation.

Validity of the assumptions in IRT

Unidimensionaility Assumption

The examination of the number and the nature of the dimensions underlying an ability test is an important aspect of construct validation (Embretson and Reise, 2000). Hattie (1985) provided a comprehensive review of 88 indices for assessing unidimensionality and concluded that many of the methods in the older psychometric literature were unsatisfactory. Methods based on nonlinear factor analysis and analyses of residuals appear to be the most promising (Hambleton et al., 1991). Unidimensionality implies that a dominant factor should

be strong enough to maintain trait level estimates unaffected by the presence of smaller specific factors (or "external "influences). Based on this, Stout (1987; 1990) proposed another procedure to judge whether a data set is unidimensional. Under the Stout framework, a test is considered unidimensional when the average between-item residual covariance after fitting a one-factor model approaches zero as the length of the test items increases.

Normally, if a unidimensional IRT model is applied, then model-fit indices implicitly evaluate how well a single dimension accounts for test data (Embretson and Reise, 2000). If a multidimensional IRT model is applied, then some compute programs (e.g., TESTFACT, NOHARM, LISCOMP) could enable researchers to develop confirmatory item-response models to assess dimensionality by developing goodness-of-fit indices, and inspect covariance residual terms after fitting models of one or more dimensions (see also Hattie, 1984).

Local Independence Assumption

A violation of local independence, called local dependence (LD), occurs when item responses depend not only on the latent trait of interest but on some other traits (Embretson and Reise, 2000). The best way of dealing with local dependence is to prevent its occurrence in the first place. Yen (1993) provided a list of suggestions for how testing situations may be arranged to minimize the chances of local dependencies occurring (e.g., construct independent items; administer test under appropriate conditions; construct separate scales). He (1984) also proposed the use of the Q3 statistic as a means of identifying pairs of test items that display local dependence. The Q3 index represents the correlation between items after partialling out the latent trait variable. Once item parameters and examinee trait levels have been estimated, it is relatively easy to compute an examinee's expected response to each test item. A residual is calculated by taking the difference between an observed raw item response and the expected item response under the IRT model. The Q3 statistic is then calculated by correlating the residual scores among item pairs. Thus, in large samples a researcher would expect Q3 to be around zero. Large positive values indicate item pairs that share some other factor and hence would violate the assumption of local independence.

Model-Data Fit

Estimating the model-data fit enables researchers to identify how close the predictions based on a specific model are to the observed data. Model-data fit assesses the extent to which the IRT model accounts for the actual test results, and helps us to understand the model-data discrepancies and their consequences. One of the most promising methods involves the analysis of item residuals. This method normally chooses an item response model, estimates corresponding item parameters and trait level, and predicts the performance of various trait level groups. The predicted results are then compared with actual results (see, e.g., Hambleton and Swaminathan, 1985; Kingston and Dorans, 1985).

Embretson and Reise (2000) proposed some specific techniques to assess the goodness-of-fit of particular IRT models. These techniques, which also require parameter estimation, are used to judge how well an IRT model represents data at the item, person, or model level (Embretson and Reise, 2000).

Assessment of item fit

There are two general approaches to evaluate the item fit (i.e., how well the IRT model explains the responses to a particular item). The first approach is the graphical procedures, which judge item fit on the basis of a comparison between an estimated ICC and an "empirical" ICC derived from actual data. Examinees are sorted by their trait level estimates and divided into several trait level groups with an equal number of examinees within each group. For a single test item, within each of the trait level groups, the "observed" percentage of item endorsements is computed. Using the within-group median trait level estimate and the proportion endorsed within a trait level group, these values are then plotted along with the estimated ICC. These plots can reveal areas along the trait level continuum where there are discrepancies between the empirical IRC and the estimated IRC.

Discrepancies (i.e., residuals) imply problems in item fit and suggest item lack of fit. The lack of fit provides diagnostic tools for researchers. Item lack of fit within a given model can be diagnosed as a problem with poor item quality, for example ambiguous wording in a multiple-choice test, and thus the item may be removed from the test form and rewritten or

replaced in future test forms. However, if a large number of poorly fitting items occur for no apparent reason, the construct validity of the test might be reconsidered and the test specifications may need to be rewritten.

The second approach to evaluate item fit is the statistical approach. Statistics that test for the significance of residuals are developed to formalize the comparison of empirical ICCs with model-based ICCs. Similar as with the graphical procedures, these statistics (e.g., Bock's chi-square index) also require estimating an ICC, scoring examinees, and grouping them into a fixed number of trait level intervals. McKinley and Mills (1985) compared the relative power of several statistical indices for assessing item fit. However, like many chi-square statistics, they found that these tests of fit are very sensitive to sample size and researchers should be careful with treating them as fixed rules of thumbs when assessing item fit.

As pointed out by Embretson and Reise (2000) there are three ways that the IRT model can represent the data. It can represent the data at the item level, person level or the model level. In the following I will discuss the various forms of fit associated with person level and model level.

Significant research has been devoted to the development of person fit indices (Meijer and Sijtsma, 1995), which attempt to estimate the validity of the IRT measurement model at the individual level (Embretson and Reise, 2000). There are many published person fit statistics, such as appropriateness measures (Levine and Rubin, 1979), caution indices (Tatsuoka, 1984; 1996), and scalability indices (Reise and Waller, 1993). Despite the different terms, all person fit indices are mainly based on the consistency of an individual's item response pattern with some proposed models of valid item responding (Embretson and Reise, 2000).

Person fit indices can detect a wide range of deviant test behaviors if such behaviors are manifested in the response pattern. However, there is no way of interpreting what the causes of lack of fit would be. Hence, it is difficult to identify the origins of a poor person fit index score (Meijer, 1996; Nering and Meijer, 1998). It seems that for a person fit statistic to function effectively, tests must be designed with the intention of detecting response aberrance. Furthermore, some person fit indices were developed within the context of specific models and are applicable only to specific IRT models, such as the Rasch model and

nonparametric models (Meijer, 1994). Even though the detection of person fit needs further development, test scores that pool person fit are normally not good predictors in practice (Embretson and Reise, 2000).

Embretson and Reise (2000) suggested that some item fit or person fit indices may be aggregated across items or examinees to obtain a general indication of model fit. In IRT models, a model can be considered appropriate to the extent to which it can reproduce the observed item responses. Given this, Thissen et al. (1986) described a procedure through which nested models can be compared by using the log-likelihood of the data given the model. In addition, Maydue-Olivares et al. (1994) presented an alternative procedure to compare the fit of competing IRT models. Their procedure can be used to compare the fit of non-nested models. Specifically, their approach is called the "ideal" observer method, and can be used to compare different parameterizations of the same data set under different IRT models. The method is based on observing the likelihoods of the item response patterns under various models and conducting a likelihood ratio test to assign the response patterns to one model versus another (see Levine et al., 1992). When the likelihoods are near equivalent across different models, it makes no difference which model is ultimately selected.

Although there have been a lot of attention devoted to IRT model selection, in reality a researcher's choices of a IRT model are somewhat more limited given his or her research goal. If a researcher has a multiple-choice test where examinees can obtain correct answers by guessing, he or she should choose the three-parameter logistic model. Also, as Embretson and Reise (2000) indicated, researchers' concern should not necessarily be on formal hypotheses testing of item or person fit, but rather on using these indices to detect the aberrant items and examinees. For example, large item fit statistics may lead the researcher to find and correct a poorly worded item or to eliminate it from the test. Identifying examinees with poor fit may allow the investigator to clean up the data set to derive better item parameter estimates.

5.3.6 Summary

As compared to the other measurement theories, Item Response Theory has its own perspective to measure a latent construct. All IRT models specify a respondent's location parameter on an underlying latent trait, which is the measure of ultimate interest. Some models also include additional parameters to better determine location and discrimination (Salzberger and Koller, 2011). A logistic s-shaped function, ICC, is typically used to model the relationship between the respondent location and the response probability, given the item properties. Through the item characteristic Curve, IRT enables researchers to detect each participant's latent trait level based on his or her item responses. Due to the invariance property, IRT is more suitable to examine the individual differences as compared to CTT and GT.

6. Contrasting three measurement theories

Classical Test Theory, Generalizability Theory, and Item Response Theory are three widely-accepted measurement theories. These theories vary in terms of how they do address measurement issues. Table 4 summarizes the major differences found between these three measurement theories. From the previous presentation we have shown that Generalizability Theory origins from Classical Test Theory and hence more commonalities can be found between GT and CTT as compared to CTT and IRT, and GT and IRT.

GT focuses on the interpretation of variance components. With its explicit focus on G-studies and D-studies it enables researchers to redesign (optimize) the measurement to achieve a higher generalizability. As compared to CTT and GT, Item Response Theory rests on different assumptions and consequently offers an alternative perspective on construct validation in that model fit (empirical versus predicted) has a central role in validating constructs. In the following, I will present some major differences among the measurement theories regarding how they measure constructs.

Different dependencies between test item and examinee

In Classical Test Theory and Generalizability Theory, examinee characteristics[28] and test characteristics cannot be separated. Each can be interpreted only in the context of the other. The true score will be defined only in terms of a particular test. The corresponding measures such as reliability (Cronbach's alpha), item-total score correlation, the standard error of measurement, are also sample dependent, implying that these measures vary across samples and especially for non-representative samples. Hence it is hard to obtain consistent reliability, item difficulty, and discrimination[29] for a single survey.

[28] In GT, examinee characteristics (the universe score) largely depends on measurement design. Meanwhile, as different measurement designs may include different facets or different levels of facet, the universe score can also been seen as dependent on the test characteristics (i.e., facets).
[29] In Classical Test Theory, the item difficulty is the proportion of correct responses, and item discrimination can be the corrected item-total correlation.

In contrast, item parameters in Item Response Theory are assumed to be not dependent on the sample used to generate the parameters, and are assumed to be invariant within a linear transformation (i.e., item information function) across divergent groups (Embretson, 1996). Furthermore, an IRT-estimated person's trait level is independent of the questions being used. The extent to how one meets this invariance property depends on how the IRT model fits the data.

Different assumption strength

Classical Test Theory relies on a set of strict and to some extent "implausible", assumptions. It requires instruments measuring the same construct to be parallel (e.g., equal means, variances, and covariance). Thus, the standard error of measurement is assumed to be the same for all examinees or participants. This is difficult, almost impossible, to accomplish given the multitude of existing surveys in marketing research. Survey responses are subject to a number of different influences, such as number of categories for each item, number of questions, order of questions, other outside measurement influences, etc.

Conversely, IRT models do not entertain the same strict assumptions. For instance, it does not require each item to be parallel. Furthermore, IRT models control for differences in item properties. Using a set of anchor items, IRT can place new items or items with different formats on a similar metric to link respondent scores. Once IRT item parameters have been estimated with an IRT model, researchers may calculate comparable scores on a given construct for respondents from that population who did not answer the same questions (Orlando, Sherbourne, and Thissen, 2000).

Generalizability theory is also based on less restrictive assumptions. Specifically, it only assumes randomly parallel tests sampled from the same universe. For instance, the sampling of persons and measurement conditions (e.g., items, occasions, interviewers, etc.) should be random. Each measurement is not required to be equivalent in mean, variance, and intercorrelation, and hence, the assumptions required by GT are less strict as compared to the ones underlying CTT.

Different ways to treat error components

Variability in measurements might be created by many different factors affected by the measurement strategy, and these factors might in turn affect the measure's quality. In Classical Test Theory, total variance in a measure's observed scores is decomposed into two components: either true score variance or error variance. Error variance is viewed as undifferentiated (Brennan, 2001; Cronbach et al., 1972). CTT cannot differentiate the effects of multiple facets, which are all pooled into a single "measurement error". Clearly, in Generalizability Theory these multiple sources of measurement error as well as the combined effects (e.g., items, occasions, raters, interactions between facets, etc.) can all be estimated. Thus, GT can be used to investigate the effects that different aspects of a measurement strategy have on the overall quality of the measure.

Different utilizations

Classical Test Theory focuses on relative (rank-order) decisions (e.g., student admission to selective colleges), whereas G theory can distinguish between relative ("norm-referenced") and absolute ("criterion-" or "domain-referenced") decisions for which a behavioral measurement is used (Rentz, 1987). Furthermore, Generalizability theory is particularly useful in developing measurement designs for subsequent studies (i.e., D study). By systematically studying various sources of error, one can develop measurement designs to reduce total error in subsequent studies.

Different reliability estimation

Statistics such as standard error of the measurement and the reliability indicate how well an instrument measures a single construct, and whether the instrument is "good". CTT yields a single estimate of reliability and standard error of measurement[30] for the measurement as a whole. They all depend on the specific sample. Furthermore, the reliability is estimated based on the correlation between true score and observed scores. Although typically just one reliability coefficient is estimated, there are several types of reliability coefficients (i.e.,

[30] The standard error measurement describes an expected score fluctuation due to error in the measurement.

internal consistency, test-retest, etc.). Sometimes researchers might find it confusing to choose among different reliability coefficients[31].

In Generalizability Theory, the definition of the generalizability coefficient (or dependability index) shares some similarities with the classical reliability coefficient. They both involve the ratio of universe (or true) score variance to observed score variance. However, there is a critical difference between them. Generalizability Theory allows multiple sources of error in observed score. The portion of variance explained by these multiple sources and their combined effects can be estimated by ANOVA. Instead of observing the correlation between universe score and observed scores, the estimation of generalizability coefficient (or dependability index) takes the multiple variance sources into account. GT develops only one coefficient. As long as the researcher is explicit about the universe to which he or she wishes to generalize, this coefficient is clear and unambiguous. Hence, in GT the problem is not to identify the "best" measure of reliability in marketing but to empirically determine what the important sources of measurement error are. Recent research have shown empirically that measurement occasions are important sources of error in marketing measures and therefore should not be ignored (Rentz, 1987).

Compared to CTT and GT, IRT utilizes a function (i.e., the information function) to estimate reliability. In Item Response Theory, measures are estimated separately for each trait level or response pattern, controlling for the characteristics (e.g., difficulty) of the items in the scale. The reliability of a test is conditional on the latent trait level. High reliability (i.e., low standard error of measurement) typically occurs in the middle of the trait continuum, and reliability is low at the low and high ends of the underlying traits (Reeve, 2002).

Different ways to increase reliability

In Classical Test Theory, items are a means to an end (DeVellis, 2003). This means they are roughly equivalent indicators of the same construct that gain strength through their aggregation. Reliability can be increased by redundancy (Kline 2000; Lee and Hooley, 2005) in that items that are more or less identical will result in high internal consistency.

[31] Even though Churchill (1979) suggested coefficient alpha absolutely should be the first measure to assess the quality of the measurement, there is still a debate in the literature.

Furthermore, based on the behavior of the most common measure of reliability, Cronbach's alpha, increasing the number of items included in the scale might also increase reliability. Guilford (1954) presented a proof that true variance increases more rapidly than error variance, if a test is lengthened by a factor of *n* parallel parts. In contrast, Generalizability Theory increases its measurement reliability through redesign. Normally, researchers would utilize the generalizable data provided by a G-study to evaluate the effectiveness of alternative. Then they could redesign (optimize) the measurement in a D-study to reduce estimated error variance and increase estimated generalizability.

Item Response Theory assesses each item's relationship to the underlying construct. Reliability is enhanced not by redundancy but by identifying better items (DeVellis, 2003). More IRT-based items typically increase the number of points along a trait level continuum (θ) that can be differentiated, but they do not increase reliability by redundancy. For example, adding more difficult questions to mathematics test extends the test's useful range upward but does not necessarily affect internal consistency. Hence, in IRT researchers can develop shorter and equally reliable scales. This is often accomplished through the use of adaptive tests that choose a set of items targeting in a respondent's level on an underlying construct (Steinberg and Thissen, 1995). Redundant items are discouraged and will actually violate the assumption of local independence.

Different validity estimation

The three theories do not differ much with regard to face validity and content validity, which is typically evaluated qualitatively. Furthermore, Item Response Theory, as well as CTT and GT, employs criterion validity to measure the correlation between its own measures and a standard variable (i.e., criterion).

In contrast, construct validity typically allows for a more rigorous investigation. In CTT fit of the data to a factor analytic model would be important evidence of construct validity, whereas in IRT the fit of the data to the particular IRT model is deemed necessary (Sinkovics and Ghauri, 2009). The model-data fit in IRT is a prerequisite for the property of invariance.

Compared to CTT and IRT, Generalizability theory does not easily fit in the traditional distinction between reliability and validity. GT assumes that the generalizability coefficient indicates how universe score can be accurately inferred from observed scores, and interprets generalizability coefficient as a validity coefficient. However, since the universe of generalization is usually not clearly defined, there still exists a debate regarding whether it is appropriate to recognize the G-coefficient as a validity coefficient.

Different object of measurement

In marketing the purpose of measurement often is not to differentiate persons but rather to differentiate products, advertisements, stores, groups of persons, etc. For instance, let us assume an internal consistency index, such as Cronbach's alpha, shows that an instrument has low internal consistency. This implies that the variance attributable to the interaction of persons and items (error variance) is large in relation to the variance attributable to persons (true score variance) so that the items do not differentiate persons well. However, if our interest is in differentiating advertisements, this index would not at all be that useful. Our interest should be in error variance relative to variance attributable to the mean advertisement ratings over persons and items. Thus classical measures of reliability are sometimes inappropriate in marketing contexts.

Conversely, Generalizability theory explicitly recognizes that persons are not always the object of measurement, and the theory is fully capable of estimating dependability for any kind of measurement objects (Rentz, 1987). Though some marketing researchers have adapted classical methods (particularly test-retest reliability) to situations in which the objects of measurement are not persons, such attempts are sometimes cumbersome. These situations could be more easily addressed with Generalizability Theory.

Item Response Theory is more suitable for testing individual difference. As the expected participants' score is computed from their responses to each item, the IRT estimated score is sensitive to differences among individual response patterns (Santor & Ramsay, 1998). This property of IRT makes the model attractive for marketing researchers investigating individual differences, but also makes it less useful for the assessment of other measurement objects.

Summary:

Item Response Theory is an alternative measurement approach to Classical Test Theory. It provides researchers greater flexibility and enables them to improve the reliability of an assessment when shorting the measurement scale. However, IRT is not a panacea for all measurement problems. Although IRT is better in terms of identifying an individual's level on a particular attribute, it is inferior in terms of scaling objects (e.g., products) as compared to CTT and especially GT. Since marketers are frequently more interested in mean scores than in individual scores there might be many situations where marketing researchers would be better served by utilizing for instance GT instead of IRT. Additionally, IRT cannot address alternative measurement models such as formative measurement. However, compared to CTT and GT, IRT is more suitable for individual ability test.

Generalizability theory is well suited for complex measurement strategies in which multiple facets might affect measurement quality. Classical Test Theory is inappropriate in some marketing situations. For simple measurement designs, Classical Test Theory is sometimes appropriate, but when the design is more complex and when the facet of differentiation is not persons, Generalizability Theory is preferable as a means of assessing and improving the reliability of marketing measures. Generalizability Theory would be even more attractive if there are multiple sources of measurement error that need to be taken into account.

Table 4: Comparison of Classical Test Theory, Generalizability Theory and Item Response Theory

	Origin	Construct Definition	Construct Measure	Error components
Classical Test Theory	Psychometrics	**1. Denotation** • Denoted as "True Score", X_T **2. "Expected Value" Definition** • A construct is defined as the expected value of observed scores under the hypothetical situation that each repeated observation is independent and parallel.	**1. Measured by mean** • $X_T = E\ (X_o)$ **2. Relation between construct and its measures** • A linear relation between construct and its observation. • The value of "true score" is dependent of the sample selected.	1. $X_E = X_O - X_T$ 2. X_E is assumed to be random. However, in reality it contain both systematic components as well as a random component.
Generalizability Theory	Psychometrics	**1. Denotation** • Denoted as "Universe Score", μ_p • Also refer to "Object of Measurement"; or "Facet of Differentiation" **2. "Expected Value" Definition** • A construct is defined as the expected value of observed scores over all observations in the universe of generalization.	**1. Measured by mean** • $\mu_p = E\ (\mu_0)$ **2. Relation between construct and its measures** • A linear relation between construct and its observation. • The value of "universe score" is dependent of the sample selected.	1. $\mu_e = \mu_0 - \mu_p$ 2. Denoted as "Facets", or "Facet of Generalization" 2. The variance of error components will be divided into different facets. The effect of each facet and the interactions between the various facets can be explicitly assessed.
Item Response Theory	Psychometrics	**1. Denotation** • Denoted as "The Trait (Ability) Level", θ **2. "Local Independence" Definition** • A construct is defined by its ability to completely explain the dependence of its measures (observed variables).	**1. Measured by a function** • Item Characteristic Curves (specific equation depends on the IRT model selected) **2. Relation between construct and its measures** • A non-linear relation between construct and its observation. • "The trait level" and item parameter is independent. Can be assessed by how well the model fits the test data.	Error components in Item Response Theory can be the residuals or discrepancies (i.e., lack of fit) between the predicted item characteristic curve and the empirical item characteristic curve.

Table 4: Comparison of Classical Test Theory, Generalizability Theory and Item Response Theory (Continue)

	Assumption Involved	Normal Process of Construct Validation
Classical Test Theory	1. Observed score is determined by respondent's true score and measurement error score. Specifically, $X_O = X_T + X_E$. 2. The measurement error is randomly distributed. 3. The random errors are not correlated with random errors of other items measuring X_T or with true score (X_T). 4. Each test form is parallel (i.e., equal means, variance, and covariance).	**1. Reliability Estimation**: • The ratio of true score variance to the expected observed score variance. • Different types of reliability: Test-retest; Parallel-forms; Internal consistency (recommended by many researchers). **2. Validity Estimation** • Different types of validity: Face Validity; Content Validity; Criterion Validity; Construct Validity (recommended by many researchers). 3. Unidimensionality should also be tested as a support for reliability.
Generalizability Theory	1. The universe of observations should contain all possible combinations of the conditions of the facets. 2. Tests are randomly parallel. (i.e., test content is assumed to be a random sample from a defined domain or universe).	**1. Generalizability Coefficient:** • Used for relative decisions • The ratio of universe score variance to the expected observed score variance. **2. Dependability Index** • Used for absolute decisions • The ratio of universe score variance to the expected observed score variance. 3. Generalizability Theory does not easily fit in the traditional distinction between reliability and validity. The focus of GT is on analyzing variance components instead of these summary coefficients.
Item Response Theory	1. The item characteristic curve reflects the true relationship among the construct and its observable variables. 2. The majority of IRT models assume "unidimensionality". However the unidimensionality assumption does not work for multidimensional item response models. 3. All IRT models assume "local independency".	**1. Reliability Estimation:** • Item information function • Standard error of measurement **2. Validity Estimation:** • Assumption Validity • Model-data fit for item level (emphasized by many researchers), person level, and model level

7. Conclusion and Recommendation

Over the years a large number of new constructs have been introduced in marketing. Furthermore, it appears that new constructs are introduced at an increasingly growing rate from year to year. Certainly these new constructs have contributed to increased explanatory power of central variables in marketing. However, it can also be questioned whether all these new constructs actually exist.

In this thesis, I have discussed and contrasted three different theories concerning measurement: Classical Test Theory, Generalizability Theory and Item Response Theory. The main research question addressed in the thesis was the following: how do these three measurement theories differ from each other in terms of underlying rationale, assessment of reliability and construct validity? To investigate this question the three psychometric measurement theories were reviewed.

Based on the review of the three measurement theories and the discussion of strength and weaknesses associated with these measurement theories the following conclusions can be made:

1. In the analysis of marketing constructs, researchers select preferred measurement theory based on various criteria. No standard or consistent selection process is found in the literature for how to pick "proper" psychometric theories. Most studies stick with traditional marketing measurement approaches which typically stem from Classical Test Theory. Typically no justification is offered as to why researchers go with a traditional approach for a specific research project. Rigdon et al. (2011) stated that marketing research is full with uncritical application of existing measurement procedures. Some researchers have introduced other psychometric models such as Generalizability Theory or Item Response Theories and claimed that models based on these theories perform better as compared to traditional models, usually referring to CTT. However, it is not possible to verify a measurement model by itself (although you can test whether some assumptions are met) since the model by itself has a more or less axiomatic status. Given that the construct is latent we can only infer the effects

of the latent constructs based on empirical observations. Hence, the adequacy of a measurement model must be assessed based on the logic of the model and the testable consequences (which sometimes are fairly limited).

2. The three measurement theories differ not only in terms of how they estimate constructs (different validity indices, procedures, involved assumptions), but also differ in terms of the importance attributed to construct validity. Given its longer history formal procedures for construct validation have already been developed within CTT framework and researchers commonly follow those procedures. Conversely, G-theory does not easily fit into the traditional validity and reliability assessment and interprets generalizability coefficient as a potential validity coefficient. However, given the reliance on the assumption of a properly defined universe, most estimated generalizability coefficients might not qualify as validity coefficients. Hence, it seems as construct validity is not receiving sufficient attention in GT. In Item Response Theory the emphasis is on testing the relevant assumptions obtained through goodness-of-fit tests and then treat the outcome of these tests as support for construct validity.

3. In most applications of these measurement theories, other types of validity (e.g., face validity or content validity) received limited attention. However, some researches show that these validities are not as unimportant as they appear to be. Poor face validity and content validity, which may not be detected statistically, would lead to a low validity of the whole research. Hence, to properly use these validities requires further studies.

4. Validating a construct includes not only testing the variations in the attribute causally produce variation in the measure, but also testing whether the concept of interest really exists (Borsboom and Mellenbergh, 2004). As Figure 1 shows, the existence of the concept is the prerequisite of the whole measurement process. However, similar to problems associated with theory selection, existence is hard to prove, but the covariance between the construct and its measures can be assessed through statistical means based on empirical observations. Starting in the late 70's there has been an extensive debate on construct validation in marketing resulting in more or less agreed upon procedures for measurement (e.g., reporting validity indices). However, all the

validity indices and model-data fit can only provide indications of validity, but cannot prove the existence of the concept. Hence, the use of statistical validation procedures are still questioned and further research is needed.

Recommendations

1. Carefully define the constructs. MacKenzie et al. (2003) concluded that a good definition should a) specify the construct's conceptual theme, b) in unambiguous terms, c) in a manner that is consistent with prior research, and that (d) clearly distinguish it from related constructs. In Generalizability theory, this definition should also include the universe being studied.

2. Develop measures that adequately represent the construct of interest. Paradigms of developing good measures have been proposed since late 1970s. However, it is worth repeating. Also, face validity and content validity should be considered.

3. The first and second criteria may be difficult to achieve simply from a statistical perspective, and cross discipline collaboration is demanded in order to ensure a valid and applicable research outcome. Although knowledge about human behavior is accumulating faster than ever before, it is still complex to model psychological attributes with number and formula. From the marketing perspective, there are sufficient reasons to believe that cooperation with computer science, neuroscience, psychology or philosophy is more than necessary.

4. The theory selection process deserves more attention. Psychometric theories should be employed to measure marketing construct in a proper manner, which may vary depending on the research problems, respondent groups, and types of scales. Construct validity must still receive sufficient attention. Also, the new tendency of coexistence of different theories, such as multifaceted-IRT etc., is an interesting development that allows for superior measurement performance.

Future Study

In this paper I have only compared measurement theories at a conceptual and theoretical level. Conclusions have been made based upon literature review and structural analysis. Future work should consider model applications at an empirical level.

As mentioned above, a fundamental question still remains and requires further research: does the concept of interest really exists? Hopefully I can look into this in working on my PhD dissertation. In the dissertation I will approach this problem by a triangulation of methods from different perspectives, such as procedures found in traditional approaches (nomological nets) and in neuroscience.

Reference

Aaker David, A. (1991). Managing brand equity. Capitalizing on the value of a brand name.

Aaker, J. L. (1997). Dimensions of brand personality. Journal of Marketing research, 347-356.

Ackerman, T. A. (1994). Using multidimensional item response theory to understand what items and tests are measuring. Applied Measurement in Education, 7(4), 255-278.

Aggarwal, P. (2004). The effects of brand relationship norms on consumer attitudes and behavior. Journal of Consumer Research, 31(1), 87-101.

Allen, M. J., & Yen, W. M. (1979). Introduction to measurement theory. Monterey, CA: Brooks/Cole.

Anderson, J. C., & Gerbing, D. W. (1982). Some methods for respecifying measurement models to obtain unidimensional construct measurement. Journal of Marketing Research, 453-460.

Andrich, D. (2004). Controversy and the Rasch model: A characteristic of incompatible paradigms?. Medical care, 42(1), I-7.

Bagozzi, R. P. (1981). Attitudes, intentions, and behavior: A test of some key hypotheses. Journal of personality and social psychology, 41(4), 607.

Bartholomew, D. J., Langeheine, R., & Tzamourani, P., (1996). Application of latent trait and latent class models in social sciences, Munster: Waxman.

Bentler, P. M. (1982). Linear systems with multiple levels and types of latent variables. In Systems Under Indirect Observation, ed. KG Joreskog, H Wold, pp. 101–30. Amsterdam: North-Holland

Baker, F. B. (2001). The basics of item response theory. Second Edition. ERIC Clearinghouse on Assessment and Evaluation.

Baker, F. B., & Kim, S. H. (2004). Item response theory: Parameter estimation techniques (Vol. 176). CRC Press.

Bollen, K. A. (1989). Structural equations with latent variables. New York: Wiley.

Bollen, K. A. (2002). Latent variables in psychology and the social sciences. Annual review of psychology, 53(1), 605-634.

Borsboom, D. (2009). Measuring the mind: Conceptual issues in contemporary psychometrics. Cambridge University Press.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. Psychological review, 111(4), 1061.

Blalock, H. M. (1968). The measurement problem: a gap between the languages of theory and research. Methodology in social research, 5-27.

Bloemer, J., & De Ruyter, K. O. (1998), "On the relationship between store image, store satisfaction and store loyalty." European Journal of Marketing 32(5/6), 499-513

Bloemer, J., De Ruyter, K. O., & Wetzels, M. (1999). Linking perceived service quality and service loyalty: a multi-dimensional perspective. European Journal of Marketing, 33(11/12), 1082-1106.

Brennan, R. L. (2001). An essay on the history and future of reliability from the perspective of replications. Journal of Educational Measurement, 38(4), 295-317.

Brennan, R.L. (2001). Generalizability Theory. Springer Verlag, New York

Breivik, E., & Thorbjørnsen, H. (2008). Consumer brand relationships: an investigation of two alternative models. Journal of the Academy of Marketing Science, 36(4), 443-472.

Brown, T. J., & Dacin, P. A. (1997). The company and the product: corporate associations and consumer product responses. The Journal of Marketing, 68-84.

Carroll, B. A., & Ahuvia, A. C. (2006). Some antecedents and outcomes of brand love. Marketing Letters, 17(2), 79-89. Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological bulletin, 56(2), 81.

Cooper, D. R., & Emory, C. W. (1995). Business Research Methods, Chicago: Richard D. Irwin.

Conger, A. J. (1981). A comparison of multi-attribute Generalizability strategies. Educational and Psychological Measurement, volume 41 issue 1 (30 November 1980), pages 121-130

Chang, P. L., & Chieng, M. H. (2006). Building consumer–brand relationship: A cross-cultural experiential view. Psychology & Marketing, 23(11), 927-959.

Chaston, I., & Baker, S. (1998). Relationship influencers: determination of affect in the provision of advisory services to SME sector firms. Journal of European Industrial Training, 22(6), 249-256.

Churchill Jr, G. A. (1979). A paradigm for developing better measures of marketing constructs. Journal of marketing research, 64-73.

Churchill Jr, G. A., & Surprenant, C. (1982). An investigation into the determinants of customer satisfaction. Journal of Marketing research, 491-504.

Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. Holt, Rinehart and Winston, 6277 Sea Harbor Drive, Orlando, FL 32887.

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963). “Theory of Generalizability: A Liberalization of Reliability Theory,” British Journal of Statistical Psychology, 16, 137-63.

Cronbach, L. J. (1971). Test validation. Educational measurement, 2, 443-507.

Cronbach, L. J., (1972). Gleser, G. C., Nanda, H., & Rajaratnam, N. The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles. New York: John Wiley & Sons, Inc.

Crosby, L. A., Evans, K. R., & Cowles, D. (1990). Relationship quality in services selling: an interpersonal influence perspective. The journal of marketing, 68-81.

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS quarterly, 319-340.

Day, George S. (1976). A two-dimensional concept of brand loyalty. In Mathematical Models in Marketing (pp. 89-89). Springer Berlin Heidelberg.

De Ayala, R. J. (2009). The theory and practice of item response theory. Guilford Press.

De Jong, M. G., Steenkamp, J. B. E., Fox, J. P., & Baumgartner, H. (2008). Using item response theory to measure extreme response style in marketing research: A global investigation. Journal of Marketing Research, 45(1), 104-115.

Denga, D. I., (2003). Educational measurement continuous assessment and psychological testing. Calabar: Rapid Educational Publishers.

DeVellis, R. F. (2003). Scale Development: Theory and Applications. Second Edition. Applied Social Research Methods Series Volume 26. Sage Publications.

Doll, W. J., & Torkzadeh, G. (1988). The measurement of end-user computing satisfaction. MIS quarterly, 259-274.

Downing, S. M. (2006). Face validity of assessments: faith-based interpretations or evidence-based science?. Medical education. 40(1).

Edward E. Rigdon, Kristopher J. Preacher, Nick Lee, Roy D. Howell, George R. Franke, Denny Borsboom, (2011) "Avoiding measurement dogma: a response to Rossiter", European Journal of Marketing, Vol. 45 Iss: 11/12, pp.1589 - 1600.

Ehrenberg, A. S. C. (1972, 1988), Repeat-Buying: Theory and Application, 2nd Edition, New York: Oxford University Press

Embretson, S. E. (1996). The new rules of measurement. Psychological Assessment, 8(4), 341.

Embretson, S. E., & Reise, S. P. (2000). Item response theory. Psychology Press.

Fournier, S. (1998). Consumers and their brands: developing relationship theory in consumer research. Journal of consumer research, 24(4), 343-353.

Flannery, W. P., Reise, S. P., & Widaman, K. F. (1995). An item response theory analysis of the general and academic scales of the Self-Description Questionnaire II. Journal of Research in Personality, 29, 168-188.

Fraley, R. C., Waller, N. G., & Brennan, K. A. (2000). An item response theory analysis of self-report measures of adult attachment. Journal of personality and social psychology, 78(2), 350.

Furr, R. M., & Bacharach, V. R. (2008). Psychometrics: An Introduction. Sage Publications.

Gerbing, D. W., & Anderson, J. C. (1988). An updated paradigm for scale development incorporating unidimensionality and its assessment. Journal of Marketing research, 186-192.

Geuens, M., Weijters, B., & De Wulf, K. (2009). A new measure of brand personality. International Journal of Research in Marketing, 26(2), 97-107.

Ghiselli, E. E., Campbell, J. P., & Zedeck, S. (1981). Measurement theory for the behavioral sciences. San Francisco: Freeman.

Glymour, C., Scheines, R., Spirtes, P., & Kelly, K. (1987). Discovering Causal Structure. Academic Press. New York.

Gonçalves, H. M. M. (2012). Multi-group invariance in a third-order factorial model: Attribute satisfaction measurement. Journal of Business Research.

Green, B. F. (1954). Attitude measurement. In G. Lindzey (Ed.), Handbook of social psychology (Vol. 1, pp. 335–369). Cambridge, MA: Addison-Wesley.

Grohmann, B. (2009). Gender dimensions of brand personality. Journal of Marketing Research, 46(1), 105-119.

Guilford, J. P. (1954). Psychometric methods. New York: McGraw-Hill.

Gummesson, E. (1996). Relationship marketing and imaginary organizations: a synthesis. European journal of Marketing, 30(2), 31-44.

Guttman, L. (1945). A basis for analyzing test-retest reliability. Psychometrika, 10, 255–282.

Hale, C. D., & Astolfi, D. (2011). Evaluating Education and Training Services: A Primer. Available at:

http://www.charlesdennishale.com/books/eets_ap2/Title%20Page%20and%20Table%20of%20Contents%202.1.pdf [Assessed: 2013-07-28]

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory. Sage Publications.

Hambleton, R. K., & Swaminathan, H. (1985). Item response theory: Principles and applications. Boston: Kluwer-Nijhoff Publishing.

Hattie, J. A., (1984). An empirical study of various indices for determining unidimensionality. Multivariate Behavioral Research, 19, 139-164

Hattie, J. (1985). Methodology review: assessing unidimensionality of tests and ltenls. Applied Psychological Measurement, 9(2), 139-164.

Holden, R. R., & Jackson, D. N. (1979). Item subtlety and face validity in personality assessment. Journal of Consulting and Clinical Psychology, 47(3), 459.

Itsuokor, D. E. (1986). Essentials of Tests and Measurements. Ilorin: Woye and Sons.

Jackson, P. H., & Agunwamba, C. C. (1977). Lower bounds for the reliability of the total score on a test composed of non-homogeneous items: I: Algebraic lower bounds. Psychometrika, 42(4), 567-578.

Jacoby, Jacob, & Kyner, David B. (1973). Brand loyalty vs. repeat purchasing behavior. Journal of Marketing research, 1-9.

Jacoby, Jacob., & Chestnut, Robert W. (1978). Brand loyalty: Measurement and management (p. 157). New York: Wiley.

Jarvis, C. B., MacKenzie, S. B., & Podsakoff, P. M. (2003). A critical review of construct indicators and measurement model misspecification in marketing and consumer research. Journal of consumer research, 30(2), 199-218.

Jones, L. V., Thissen, D. (2007). A history and overview of psychometrics. Handbook of statistics, 26, 1-27.

Judd, C. M., Smith, E. R., & Kidder, L. H. (1991). Research methods in social science. Fort Worth: Holt, Rinehart, and Wanston.

Kane, M. T. (1992). An argument-based approach to validity. Psychological Bulletin; Psychological Bulletin, 112(3), 527.

Kaplan, Abraham. (1964). The Conduct of Inquiry: Methodology for Behavioral Science. Scranton, PA: Chandler Publishing Co.

Keller, K. L. (1993). Conceptualizing, measuring, and managing customer-based brand equity. The Journal of Marketing, 1-22.

Kelley, T. L. (1927). Interpretation of educational measurements. Yonkers: World Book.

Kim, J. H., Ritchie, J. B., & McCormick, B. (2012). Development of a scale to measure memorable tourism experiences. Journal of Travel Research, 51(1), 12-25.

Kingston, N. M., & Dorans, N. J. (1985). The analysis of item-ability regressions: An exploratory IRTmodel fit tool. Applied Psychological Measurement, 281-288.

Klein, D. F., & Cleary, T. (1967). Platonic true scores and error in psychiatric rating scales. Psychological Bulletin, 68(2), 77.

Kline, P. (2000). A psychometrics primer. Free Assn Books.

Knox, S., & Walker, D. (2001). Measuring and managing brand loyalty. Journal of Strategic Marketing, 9(2), 111-128.

Lee, N., & Hooley, G. (2005). The evolution of "classical mythology" within marketing measure development. European Journal of Marketing, 39(3/4), 365-385.

Levine, M. V., & Rubin, D. B. (1979). Measuring the appropriateness of multiple-choice test scores. Journal of Educational and Behavioral Statistics, 4(4), 269-290.

Levine, M. V., Drasgow, F., Williams, B., McCusker, C., & Thomasson, G. L. (1992). Distinguishing between item response theory models. Applied Psychological Measurement, 16.

Likert, R. (1931). A technique for the measurement of attitudes. Archives of Psychology. New York: Columbia University Press.

Lord, F. M., Novick, M. R., & Birnbaum, A. (1968). Statistical theories of mental test scores. Oxford, England: Addison-Wesley.

Lord, F. M. (1974). Quick estimates of the relative efficiency of two tests as a function of ability level. Journal of Educational Measurement, 11(4), 247-254.

Lord, F. M. (1980). Applications of item response to theory to practical testing problems. Lawrence Erlbaum.

MacKenzie, S. B., Podsakoff, P. M., & Podsakoff, N. P. (2011). Construct measurement and validation procedures in MIS and behavioral research: integrating new and existing techniques. MIS quarterly, 35(2), 293-334.

Masters, G. N., & Wright, B. D. (1984). The essential process in a family of measurement models. Psychometrika, 49(4), 529-544.

Masters, G. N., Wright, B. D., W. J. van der Linden, R. K. Hambleton Masters, G. N. , & Wright, B. D. (1997). The partial credit model. In W. J. van der Linden & R. K. Hambleton (Eds.), Handbook of modern item response theory (pp. 101–121). New York: Springer.

Maydeu-Olivares, A., Drasgow, F., & Mead, A. D. (1994). Distinguishing among parametric item response models for polychotomous ordered data. Applied Psychological Measurement, 18.

McDonald, R. P. (1989). Future directions for item response theory. International journal of educational research, 13(2), 205-220.

McDonald, R. P. (1999). Test theory: A unified treatment. Mahwah, NJ: Erlbaum.

McKinley, R. L., & Mills, C. N. (1985). A comparison of several goodness-of-fit statistics. Applied Psychological Measurement, 9, 49-57.

McMullan, R., & Gilmore, A. (2003). The conceptual development of customer loyalty measurement: a proposed scale. Journal of Targeting, Measurement and Analysis for Marketing, 11(3), 230-243.

McMullan, R., & Gilmore, A. (2008). Customer loyalty: an empirical study. European Journal of Marketing, 42(9/10), 1084-1094.

Mehrens, W. A., & Lehmann, I.J. (1984). Measurement and Evaluation in Education and Psychology. 3rd ed. New York: Holt, Rinehart and Winston

Meijer, R. R., & Sijtsma, K. (1995). Detection of aberrant item score patterns: A review of recent developments. Applied Measurement in Education, 8(3), 261-272.

Meijer, R. R. (1996). Person-fit research: An introduction. Applied Measurement in Education, 9(1), 3-8.

Meijer, R. R. (1994). The number of Guttman errors as a simple and powerful person-fit statistic. Applied Psychological Measurement, 18(4), 311-314.

Mellenbergh, G. J. (1996). Measurement precision in test score and item response models. Psychological Methods, 1(3), 293.

Mellenbergh, G. J. (1994). A unidimensional latent trait model for continuous item responses. Multivariate Behavioral Research, 29(3), 223-236.

Menon, A., & Varadarajan, P. R. (1992). A model of marketing knowledge use within firms. The Journal of Marketing, 53-71.

Meyer, John. P., & Allen, Natalie J. (1991). A three-component conceptualization of organizational commitment. Human resource management review, 1(1), 61-89.

Morgan, R. M., & Hunt, S. D. (1994). The commitment-trust theory of relationship marketing. the journal of marketing, 20-38.

Nering, M. L., & Meijer, R. R. (1998). A comparison of the person response function and the lz person-fit statistic. Applied Psychological Measurement, 22(1), 53-69.

Neuman, W. L. (2007). Basics of Social Research: Quantitative and Qualitative Approaches (2nd ed.). Boston: Allyn and Bacon.

Novak, T. P., Hoffman, D. L., & Yung, Y. F. (2000). Measuring the customer experience in online environments: A structural modeling approach. Marketing science, 19(1), 22-42.

Nunnally, Jum C. (1978). Psychometric Theory. New York: McGraw-Hill Book Company, Second Edition.

Oliver, R. L. (1999). Whence consumer loyalty?. the Journal of Marketing, 33-44.

Orlando, M., Sherbourne, C. D., & Thissen, D. (2000). Summed-score linking using item response theory: Application to depression measurement. Psychological Assessment, 12(3), 354.

Parasuraman, A., Zeithaml, V. A., & Berry, L. L. (1985). A conceptual model of service quality and its implications for future research. The Journal of Marketing, 41-50.

Park, C.W., Jaworski, B.J., MacInnis, D.J. (1986), "Strategic brand concept image management", Journal of Marketing, Vol. 50 pp.135-45.

Peattie, K., & Charter, M. (1994). Green marketing. The marketing book, 5, 726-755.

Percy, L., & Rossiter, J. R. (1992). A model of brand awareness and brand attitude advertising strategies. Psychology & Marketing, 9(4), 263-274.

Peter, J. P. (1979). Reliability: a review of psychometric basics and recent marketing practices. Journal of marketing research, 6-17.

Peter, J. P. (1981). Construct validity: a review of basic issues and marketing practices. Journal of Marketing Research, 133-145.

Punniyamoorthy, M., & Raj, M. P. M. (2007). An empirical model for brand loyalty measurement. Journal of Targeting, Measurement and Analysis for Marketing, 15(4), 222-233.

Rasch, G. (1960). Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests. Copenhagen, Danmarks pædagogiske Institut.

Reckase, M. D. (1997). The past and future of multidimensional item response theory. Applied Psychological Measurement, 21(1), 25-36.

Reeve, B. B. (2002). An introduction to modern measurement theory. National Cancer Institute.

Reise, S. P., & Waller, N. G. (1993). Traitedness and the assessment of response pattern scalability. Journal of Personality and Social Psychology, 65(1), 143.

Rentz, J. O. (1987). Generalizability theory: A comprehensive method for assessing and improving the dependability of marketing measures. Journal of Marketing Research, 19-28.

Roberts, J. S., Donoghue, J. R., & Laughlin, J. E. (2000). A general item response theory model for unfolding unidimensional polytomous responses. Applied Psychological Measurement, 24(1), 3-32.

Richins, M. L. (1983). Negative word-of-mouth by dissatisfied consumers: a pilot study. The Journal of Marketing, 68-78.

Rozeboom, W. W. (1978). Estimation of cross-validated multiple correlation: A clarification. Psychological Bulletin, 85(6), 1348.

Rudolf Sinkovics, & Ghauri (Eds.). (2009). New challenges to international marketing (Vol. 20). Emerald Group Publishing.

Salzberger, T., & Koller, M. (2012). Towards a new paradigm of measurement in marketing. Journal of Business Research.

Schmidt, F. L., & Hunter, J. E. (1999). Theory testing and measurement error. Intelligence, 27(3), 183-198.

Schmitt, B., Zarantonello, L., & Brakus, J. (2009). Brand experience: what is it? How is it measured? Does it affect loyalty?. Journal of Marketing, 73(3), 52-68.

Selnes, F., & Gønhaug, K. (2000). Effects of supplier reliability and benevolence in business marketing. Journal of Business Research, 49(3), 259-271.

Singh, J. (2004). Tackling measurement problems with Item Response Theory: Principles, characteristics, and assessment, with an illustrative example. Journal of Business Research, 57(2), 184-208.

Shavelson R J, Webb N M, & Rowley G L. (1989). Generalizability theory. American Psychologist, 1989, 44(6): 922.

Shavelson, R. J., & Webb, N. M. (1991). Generalizability Theory. Sage Publications.

Singhapakdi, A., & Vitell, S. J. (1990). Marketing ethics: Factors influencing perceptions of ethical problems and alternatives. Journal of Macromarketing, 10(1), 4-18.

Song, X. M., Xie, J., & Dyer, B. (2000). Antecedents and consequences of marketing managers' conflict-handling behaviors. The Journal of Marketing, 50-66.

Spray, J. A., Davey, T. C., Reckase, M. D., Ackerman, T. A., & Carlson, J. E. (1990). Comparison of two logistic multidimensional item response theory models (No. ACT-RR-ONR-90-8). American Coll Testing Program Iowa City IA.

Steinberg, L., & Thissen, D. (1995). Item response theory in personality research. Personality research, methods, and theory: A festschrift honoring Donald W. Fiske, 161-181.

Stevens, S. S. (1946). On the theory of scales of measurement. Science, vol. 103, No. 2684

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. Psychometrika, 52(4), 589-617.

Stout, W. (1990). A new item response theory modelling approach and applications to unidimensionality assessment and ability estimation. Psychometrika, 55, 293-325.

Tatsuoka, K. K. (1984). Caution indices based on item response theory. Psychometrika, 49(1), 95-110.

Tatsuoka, K. K. (1986). Diagnosing cognitive errors: Statistical pattern classification based on item response theory. Behaviormetrika, 19, 73-86.

Torgerson, W. S. (1958). Theory and methods of scaling. New York: Wiley.

Torgerson, W. S. (1962). Theory and methods of scaling. Second edition New York: Wiley.

Traub, R. (1997). Classical Test Theory in Historical Perspective. Educational Measurement: Issues and Practice, 16 (4)

Trochim, W. (2000). The Research Methods Knowledge Base, 2nd Edition. Atomic Dog Publishing, Cincinnati, OH.

Thissen, D., Steinberg, L., & Gerrard, M. (1986). Beyond group-mean differences: The concept of item bias. Psychological Bulletin, 99, 118-128.

Thorndike, E. L. (1918). Educational psychology: Briefer course. Teachers College, Columbia University.

Thurstone, L. L. (1931). The measurement of social attitudes. The Journal of Abnormal and Social Psychology, 26(3), 249.

van der Linden, W. J., & Hambleton, R. K. (Eds.). (1997). Handbook of modern item response theory. Springer.

Velázquez, B. M., Saura, I. G., & Molina, M. E. R. (2011). Conceptualizing and measuring loyalty: Towards a conceptual model of tourist loyalty antecedents. Journal of Vacation Marketing, 17(1), 65-81.

Walters, D., & Lancaster, G. (1999). Value-based marketing and its usefulness to customers. Management Decision, 37(9), 697-708.

Webb, N. M. (1991). Generalizability theory: A primer. Sage Publications.

Webb, N. M., Shavelson, R. J., & Haertel, E. H. (2006). Reliability coefficients and generalizability theory. Handbook of statistics, 26, 81-124.

Weiner, I. B., & Craighead, W. E. (2010). The Corsini encyclopedia of psychology (Vol. 4). Wiley. com.

Weiss, D. J., & Davison, M. L. (1981). Test theory and methods. Annual Review of Psychology, 32(1), 629-658.

Wright, B. D. (1997). A history of social science measurement. Educational Measurement: Issues and Practice, 16(4), 33-45.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. Applied Psychological Measurement, 8, 125145.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. Journal of Educational Measurement, 30, 187213.

Appendix

Figure 11[32]: Number of literatures that adopt latent construct from selected journals

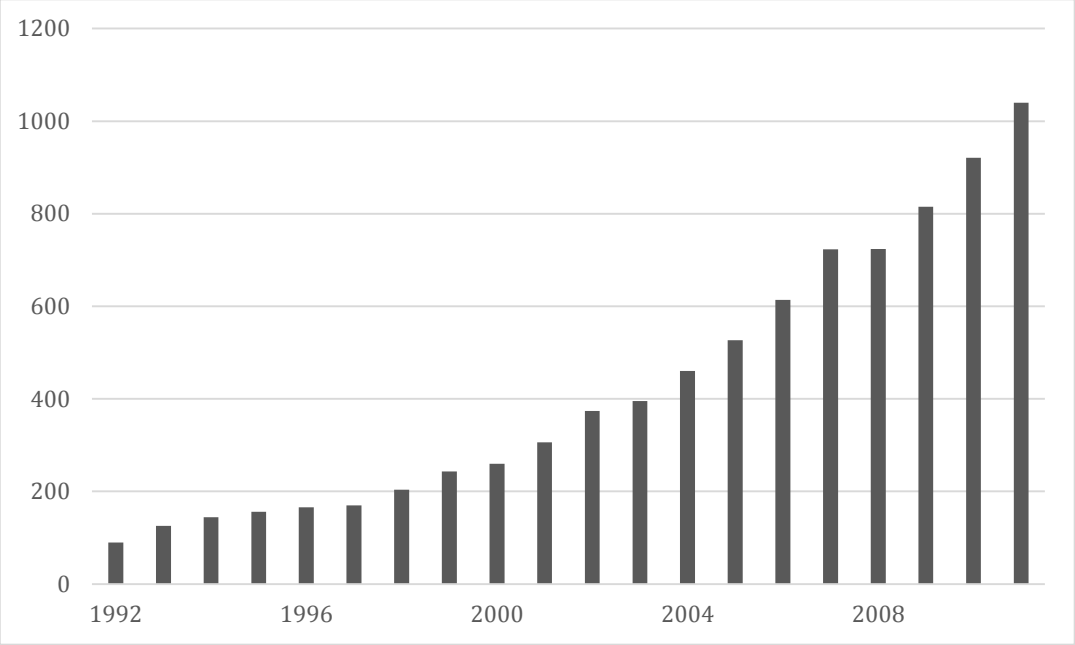
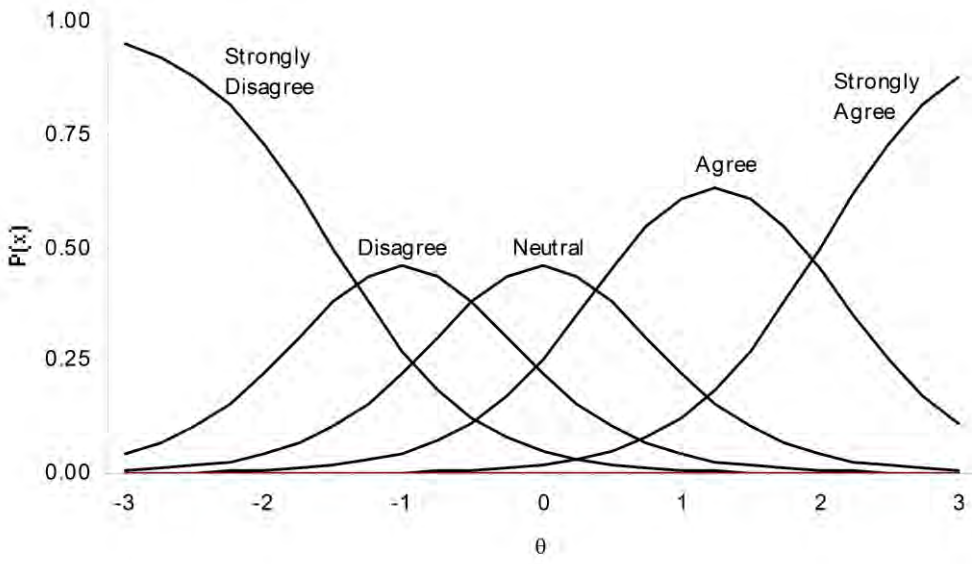


Figure 12: Category response curves for five-category item under graded response model



[32] Search criteria: using the three keywords (i.e., "construct", "latent variable", "factor analysis"), I filtered the number of literatures which mainly focus on either new construct or its measurement.

Table 5: Search result[33] for marketing literature that utilized Generalizability Theory to measure the construct empirically

Before (and include) 2000	After 2000
Peter, J 1977	Finn, A 2001
Prakash, V, & Lounsbury, J 1983	Hillebrand, B, Kok, R, & Biemans, W 2001
Steenkamp, J, & Baumgartner, H 1998	Ronald E., G 2001
Goodwin, L, Sands, D, & Kozleski, E 1991	Lin, L 2003
Rentz, JO 1987	Beverland, M, & Lockshin, L 2004
Hughes, M, & Garrett, D 1990	Finn, A 2004
HUGHES, M, & GARRETT, D 1990	Malthouse, E, Oakley, J, Calder, B, & Iacobucci, D 2004
Marcoulides, G, & Goldstein, Z 1992	Finna, A, & Kayande, U 2005
Finn, A, & Kayande, U 1997	Kim, S, & Pridemore, W 2005
Bell, D, Chiang, J, & Padmanabhan, V 1999	Steenkamp, JM 2005
Muncy, J, & Gomes, R 1992	Winkelman, W, Leonard, K, & Rossos, P 2005
Finn, A, & Kayandé, U 1999	Durvasula, S, Netemeyer, R, Andrews, J, & Lysonski, S 2006
	Peng, M, Zhou, Y, & York, A 2006
	Brush, G, & Rexha, N 2007
	Ad de, J, Wetzels, M, & Ko de, R 2008
	Guo, L, & Xiangyu, M 2008
	Eisend, M 2009
	Sharma, S, & Durvasula, S 2009
	Wilson, R, & Amine, L 2009
	Audia, P, & Rider, C 2010
	Peng, L, Cui, G, & Chunyu, L 2012
	Wang, L, & Finn, A 2012
	Wang, L 2012
	Ingenbleek, P, Tessema, W, & van Trijp, H 2013

[33] Searched in EBSCO business source premier database

Table 6: Search result for marketing literature that utilized item response theory to measure the construct empirically

Before (and include 2000)	After 2000
Bechtel, G, Ofir, C, & Ventura, J 1990	Owens, J 2001
Dodd, B, De Ayala, R, & Koch, W 1995	Singh, J 2004
Balasubramanian, S, & Kamakura Wagner, A 1989	Houran, J, Lange, R, Rentfrow, P, & Bruckner, K 2004
BALASUBRAMANIAN, S, & KAMAKURA, W 1989	Raajpoot, N 2004
Yamaguchi, J 2000	Swaminathan, S, & Bawa, K 2005
Maydeu-Olivares, A 1998	Baranowski, T, et al. 2006
	Reise, S, Meijer, R, Ainsworth, A, Morales, L, & Hays, R 2006
	Chakravarty, E, Bjorner, J, & Fries, J 2007
	Kristjansson, et al. 2007
	Baranowski, T, et al. 2008
	de Jong, M, Steenkamp, J, & Veldkamp, B 2008
	de Jong, M, Steenkamp, J, & Veldkamp, B 2009
	LaHuis, D, & Copeland, D 2009
	Salzberger, T, Holzmüller, H, & Souchon, A 2009
	Sharma, S, & Durvasula, S 2009
	Bennett, C, et al. 2010
	Rodriguez, H, & Crane, P 2011
	Rim, H, Turner, B, Betz, N, & Nygren, T 2011
	Amtmann, D, Cook, K, Johnson, K, & Cella, D 2011
	Hasford, J, & Bradley, K 2011
	Wang, L, & Finn, A 2012
	Wang, L 2012
	Coromina, L 2013