SCALE VALIDITY IN EXPLORATORY STAGES OF RESEARCH

PhD Patricea Elena BERTEA
Romanian Academy Iaşi, Romania
Email: patricia.bertea@yahoo.com
Professor PhD Adriana ZAIȚ
University "A. I. Cuza", Iaşi, Romania
Email: azait@uaic.ro

Abstract:
Scale development assumes that certain steps are to be taken in order to obtain a valid measurement instrument. Most of the researchers jump to the confirmatory stage and avoid exploratory measures. However, exploratory methods that are used in the first stages of scale development are recommended so as to avoid further problems regarding the validity of the scale. Before conducting reliability analysis and factorial analysis, exploratory methods can be applied. The main purpose of this paper is to draw the attention on alternative methods for scale validation that should be used in the exploratory phase. The role of these methods is to improve validity of results of the further confirmatory phases of research. The Lawshe (1975) content validity ratio and the Q-sorting procedure for testing construct validity are applied in the process of developing a scale for perceived risk.

Keywords: scale development, content validity, q-sorting, perceived risk

Introduction

The main purpose of this paper is to draw the attention on alternative methods for scale validation that should be used in the exploratory phase. The role of these methods is to improve validity of results of the further confirmatory phases of research. The methods are exemplified on a scale that aims to measure perceived risk in e-commerce.

Scale development has become an important research area since several seminal works (Cronbach, 1951; Nunnaly, 1967; Churchill, 1979). The use of scales in Management and Marketing research has become common since both fields deal with studies on latent variables. Thus, the methodology from Psychology is now successfully employed by researchers from the previously mentioned areas.

An important aspect in scale development is assessing validity. Validity refers to the ability of a construct to measure what it was supposed to measure (Goodwin, 2009). When assessing the validity of a scale we are actually looking how accurate the scale is (Groth-Marnat, 2009). Establishing the validity of a scale is rather difficult, especially when we are dealing with psychological variables. The main issue is that such variables are not observable and the researcher has to identify the underlying latent variables by constructing measurement instruments. Validation of measurement instruments assumes that the inferences and conclusions that are drawn in a research are actually valid (Schultz & Whitney, 2004).

Another issue when talking about validity is that it should not be confounded with reliability. Reliability, which is usually measured using the

Cronbach alpha coefficient, refers to the consistency of the measurement. A more clarifying perspective is given by Campbell and Fiske (1959), who explain that reliability is the agreement of two attempts to measure the same underlying construct through similar methods, while validity refers to the same issue, but the methods used are totally different. Cronbach alpha measures a certain type of reliability which is defined as internal consistency and offers information on how items that form a scale correlate with each other. An accepted level of internal consistency has to be at least of 0,7, but not higher than 0,9 (Cronbach, 1951), which indicates that some items might be redundant inside the scale. Alwin (2007) considers that alpha Cronbach should be used more as an internal consistency measure that shows how "a set items hangs together to form a scale" and that other approaches should be employed in assessing reliability. Among these, Alwin (2007) talks about using multi-trait multi-method/ confirmatory factor analysis to measure reliability. As far as validity is concerned, Alwin (2007) explains that "a reliable measure is not necessarily a valid one".

Types of validity

There are different types of validity that researchers should look into when developing a scale. Specialists talk about three types of validity: criterion validity, content validity and construct validity.

Criterion validity

Criterion validity stands for how well an instrument measures a variable in comparison with another instrument or a predictor. There are two types of criterion validity: concurrent and predictive validity.

Concurrent validity assumes there is another construct that measures the same variable, a construct considered to be a benchmark in the research domain. To have concurrent validity for a construct it is compulsory that there is a high correlation with the benchmark construct. Researchers can also choose the benchmark as being a totally opposed variable and in this case low correlation is expected in order to have good concurrent validity. Usually, to test for concurrent validity researchers apply two different instruments measuring the same variable on the same sample, just that one of the instruments must be a standard in the domain, with previously tested psychometric characteristics.

Predictive validity refers to the ability of a measurement instrument to predict future attitudes or behaviors. Establishing predictive validity means that data is collected twice at different moments in time, so as to check if the scale predicted or not a certain event. In this case there is also need to do a correlation between the variable we are trying to measure and another variable that is used as a criterion.

Content validity

Content validity refers to a correct definition of the domain of the latent variable that one intends to measure. Another important aspect is the identification of possible facets of the construct. Thus, when we want to measure a latent variable is important to introduce in the construct all possible items which could capture the essence of the variable (Haynes, et al., 1995). For instance, if we include items that have no connection with the variable that we generate measurement errors, while if we exclude items that we will have exclusion errors (Straub, et al., 2004).

Content validity assumes two stages (Lynn, 1986): the development stage and the judgement-quantification stage. The first stage implies the use of qualitative methods such as interviews, focus groups and, of course, an intensive review of literature. The second stage, which is intended to

quantify the validity of a scale, requires that a panel of experts evaluate the scale's items accordingly to the

Although methods have been developed for the second stage, most researchers appeal to literature review and other qualitative methods to assure content validity of the scale. This qualitative type of validation is more or less prone to subjective influences coming from the researchers. Yet, this approach is intensively used and there are few who reach for alternative quantitative methods. Nevertheless, using a more empirical method with a quantitative foundation adds more scientific value to our research and prevents validation problems to further affect results.

Content validity measures

There are several ways to test content validity using a quantitative approach.

Lawshe (1975) developed a quantitative measure for assessing content validity called the content validity ratio (CVR). The content validity ratio offers information about item-level validity. The procedure consists in using a panel of experts to rate items according to the relevance for the domain of the scale. Each item of a scale is rated on 3-point rating system (1- item is irrelevant, 2 – item is important, but not essential, 3 – item is essential). For each item a CVR is computed, that is basically the proportion of experts that considered the items important or essential for the content of the scale. There is also the possibility of having an overall measure for the content validity of the scale. This is called an index and it is computed as a mean of items' CVR values.

Another quantitative measure was proposed by Waltz & Bausell (1983) and it is called the Content Validity Index (CVI). The difference between this measure and the previous (Lawshe, 1975) is that experts rate items on a 4-points rating scale with slightly different anchors (1 – not relevant, 2 – somewhat relevant, 3 – quite relevant and 4 – very relevant). The index computation is actually a percentage given by the number of experts that rate quite relevant or very relevant an item. A total index per scale can also be computed. According to Waltz et. al (2010) the CVI per scale is recommended when there are only two experts involved in the judgment stage. When more than two judges are involved, Waltz et. al (2010) recommend to use alpha coefficient, that quantifies the extent to which there is agreement between experts.

Construct validity

Construct validity refers more to the measurement of the variable. The issue is that the items chosen to build up a construct interact in such manner that allows the researcher to capture the essence of the latent variable that has to be measured. Content validity must be assessed priori to construct validity. Construct validity implies the use of more quantitatively oriented analyses.

It is important to make the distinction between internal validity and construct validity. The first one refers to assuring a methodology that enables the research to rule out alternative explanations for the dependent variables, while construct validity is more concerned with the choice of the instrument and its ability to capture the latent variable. Internal validity becomes a problem in experimental studies, where each experimental group has to follow the same methodology in order to be able to correctly isolate the effect.

Construct validity has three components: convergent, discriminant and nomological validity. Convergent validity and discriminant validity refers to the way the construct relates to other constructs. Convergent validity tests if the items of a scale correlate higher among them and have significant higher loadings. Convergent validity can also be assessed buy checking the correlation between the instrument and

other instruments that mean to measure the same latent variable. Discriminant validity assumes that items should correlate higher among them than they correlate with other items from other constructs that are theoretically supposed not to correlate. Nomological validity tests if the construct has the same relationships with other variables that have been previously tested and confirmed in other studies.

Construct validity can be tested during early stages of research using the Q-sorting procedure. The main idea of the analysis is to separate items in construct according to their specific domain. The procedure is more close to measuring discriminant validity. There are two ways that it can be done (Storey, et al., 1997):

• Exploratory, when respondents are given the items and asked to group and identify category labels for each group of items.

• Confirmatory, when the categories are already labeled and respondents are asked to classify each item in one category.

Q-sorting is applied on experts and other persons of interest for the research. It helps eliminate items that do not discriminate well between categories.

Research methodology

The present study presents to alternative methods for assessing scale validaty: the content validity ratio and the Q-sorting procedure. Both procedures were applied on a scale that measures perceived risk in e-commerce.

For building up the construct for perceived risk in e-commerce we followed the methodology used by Jacoby and Kaplan (1972). They divided perceived risk into six dimensions: financial, performance, time, social, psychological and physical. We did not use the same dimensions as listed above, since Jacoby and Kaplan (1972) did research on products.

We aimed to study perceived risk of Internet as an alternative shopping channel. As a consequence, there was need to restate the dimensions. In order to do, that we investigated the work of Featherman and Pavlou (2003) together with Crespo, et al. (2009). In the end we defined six dimensions of perceived risk in e-commerce: financial, security/privacy, psychological, social, time/delivery and product risk. Each dimension was identified through a number of items ranging from 3 to 8, which were extracted from the literature review and in-depth interviews (table 1).

Table 1

Dimensions of perceived risk in e-commerce

Type of risk	Items
Product risk	I believe that online shopping is risky because I cannot examine the product.
	If I choose to buy online I do not have the certainty that the product will be of good quality.
	I believe online shopping is risky because I cannot touch the product before buy it.
	I cannot be sure that a product bought online has the characteristics advertised on the website.
	I believe that a product bought online will not perform as well as one bought from a bricks and mortar store.
	If I buy a product online I risk not to be given the guaranty.

Financial risk	I do not trust online payment.
	When I pay online there is an increased probability to lose the money on my credit card.
	Using online payment there is a chance I pay more due to hidden fees.
	There is a low probability to lose money for a product ordered on the Internet if I pay on delivery.
	I believe that paying by credit card is a secure payment method.
	There are high chances of losing money when paying online for a product.
	Online shopping means potential money loss due to possible Internet frauds.
	The risk of losing money when buying online is the same whether I pay by credit card or on delivery.
Security/ privacy risk	If I buy online there is a high risk that my personal data would be used without my consent.
	There is high chance that hackers take over my personal account from a e-shop.
	If I decide to buy products online I risk losing control over my personal data.
Time/ delivery risk	If I do my shopping online, there is a high risk that I receive a different product that the one I ordered.
	When I buy online I am sure that I will receive exactly the product I ordered.
	If I buy online there are low chances that my product would have a delivery delay.
	When I buy online, I not sure that the e-shop will respect the promised deadline.
Social Risk	There is small chance that my friends will change their opinion about me because of me using Internet to do shopping.
	If I buy online I am taking the risk that my friends will change their opinion about me.
	Online shopping is positively seen by my family.
	My friends do not approve online shopping.
Psychological risk	Online shopping does not fit my self-image.
	Online shopping is not compatible to my self-image.
	Online shopping gives me a state of stress because it does not fit with my self-image.
	Online shopping fits me well.

In order to apply the two methods we had to do two separate studies for which we developed two questionnaires.

Methodology for the content validity ratio

For the content validity ratio we followed the methodology explained by Lawshe (1975). We introduced all the items grouped for each type of risk. We interviewed six experts that were asked to answer if each item was "1= Irrelevant, 2=Important, but not essential and 3=Essential" for measuring a certain type of perceived risk.

Table 2

CVR questionnaire example

Product risk item	Irrelevant	Important, but not essential	Essential
I believe that a product bought online will not perform as well as one bought from a bricks and mortar store.			

Methodology for the Q-sorting procedure

For the Q-sorting study we developed a questionnaire were we included all items measuring perceived risk without showing which item belongs to which type of perceived risk. Respondents had to classify items into 6 categories: social, psychological, financial, security, product and delivery risk (table 3).

Table 3

Q-sorting questionnaire example

Risk Item	Risk Type
Online shopping gives me a state of stress because it does not fit with my self-image.	Social
	Financial
	Psychological
	Security
	Delivery
	Product

As a quantitative indicator of the Q-sorting procedure we used the correct classification percent, which describes the percent of respondents that have correctly classified an item (Straub, et al., 2004).

Results

Content Validity Ratio

To calculate the content validity ratio we used the methodology described by Lawshe (1975), which indicates that all items should be analyzed by a group of experts, each expert having the possibility to describe the item as: 1= Irrelevant, 2=Important, but not essential and 3=Essential. The formula to calculate the ratio is:

$$CVR = \frac{n - I}{N},$$

Where n – number of experts who considered the item to be “Essential” or “Important, but not essential”;

I – number of experts who considered the item “Irrelevant”;

N – total number of experts;

The logic behind the formula is that the more experts are in favor of one item as being important or essential, the more we can consider that item as being part of the construct. Thus, we can attain content validity of the construct. As one can easily see, the formula gives a negative result when less than 50% of the experts rate the item as essential or important but not essential or a null result when 50% rate it as irrelevant.

A panel formed by six experts rated the items according to Lawshe (1975) specifications. After analyzing the data, we identified 7 items which presented serious problems, CVR value being negative, which suggests that more than 50% of experts found the items to be irrelevant (table 4).

Table 4

CVR values

Item	CVR values
Product risk – I believe that a product bought online will not perform as well as one bought from a bricks and mortar store.	-0.67
Social risk – There is small chance that my friends will change their opinion about me because of me using Internet to do shopping.	-0.67
Social risk – If I buy online I am taking the risk that my friends will change their opinion about me.	-0.67
Psychological risk – Online shopping does not fit my self-image.	-0.67
Psychological risk – Online shopping gives me a state of stress because it does not fit with my self-image.	-0.33
Psychological risk – Online shopping does not fit with my self-image.	-0.67
Psychological risk – Online shopping suits my self-image.	-0.33

These results suggest that the 7 items should be removed from the construct before advancing the research.

Q-sorting

In order to calculate the percent of correct classification, we identified the frequency of respondents that checked the correct category for each item. We had items that obtained a 100% correct classification – 3 items, items that had percents higher than 70% – 22 items, but also items with lower percents -4 items. We considered items with a low classification percent those who were below 60% (table 5).

Taking into account that more than 80% of all 26 items were correctly classified, we can consider that the scale has a good level of discriminant validity. However, it is important to further analyze those items that were not correctly recognized as belonging to a certain category of risk.

Table 5

Q-sorting results (items with low classification)

Risk type	Item	Percent
Psychological	Online shopping does not fit my self-image.	0.52
Security/ privacy	There is high chance that hackers take over my personal account from a e-shop.	0.59
Time/delivery	If I do my shopping online, there is a high risk that I receive a different product that the one I ordered.	0.52
	When I buy online I am sure that I will receive exactly the product I ordered.	0.22

Conclusions

There is only one item that presented common problems in both procedures, the one belonging to psychological risk. However, the objective of the research was no to see whether there are items with problems in both cases, but to identify items that do not match validity. So, CVR was measured to test for content validity, while Q-sorting was applied to test for construct validity, more specifically discriminant validity of items.

Both alternative methods revealed items with significant problems, items that should be removed in next stages of the study or should be refined in order to express more clearly a certain type of risk.

The major implications of this research rest in the importance of correctly developing a measurement instrument for a latent variable. There is need for applying alternative methods to test scale validity especially when we develop a whole new construct and we use qualitative methods such as in-depth interviews or focus groups, but also when we want to use a scale that was previously developed, but never used on a certain sample. The concern for applying these types of methods should exist whenever the aim is to raise the quality of a research. That would show a profound investigation of all possible issues which may affect scale validity.

Further research should concentrate on establishing how these methods can improve convergent validity, discriminant validity and nomological validity. Moreover, it could be useful to examine who are the most appropriate respondents for each method. If we have to use only experts or we could also use non-experts, just consumers. An interesting approach would be to compare results coming from two different samples and to see whether respondents' type is an issue. The problem is, however, that the experts sample will always be smaller than the consumers' one and it is difficult to obtain representativity.

The value of this research stands in the revival of rather isolated methods of scale validation that can prove high utility in exploratory phases of research. Content validity ratio and Q-sorting are less employed, so we wanted to introduce them and raise researchers' interest for these alternative methods.

REFERENCES

Alwin, D. (2007), 'Margins of error: A study of reliability in survey measurement', 547.

Campbell, D. & Fiske, D. (1998), 'Convergent and discriminant validation by the multitrait-multimethods matrix', Personality 56, 162.

Churchill Jr., G. A. (1979), 'A Paradigm for Developing Better Measures of Marketing Constructs.', Journal of Marketing Research (JMR) 16(1), 64 - 73.

Crespo, Á.. H., del Bosque, I. R. & de los Salmones Sánchez, M. M. G. (2009), „The Influence Of Perceived Risk On Internet Shopping Behavior: A Multidimensional Perspective", Journal of Risk Research, 12(2), 259–277.

Cronbach, L. (1951), „Coefficient alpha and the internal structure of tests", Psychometrika 16(3), 297-334.

Featherman, M. S. & Pavlou, P. A. (2003), „Predicting e-services adoption: a perceived risk facets perspective", International Journal of Human-Computer Studies, 59(4), 451 - 474.

Goodwin, C. (2009), Research in psychology: Methods and design, Wiley.

Groth-Marnat, G. (2009), Handbook of psychological assessment, Wiley.

Gwet, K. (2001), Handbook of inter-rater reliability.

Haynes, S.; Richard, D. & Kubany, E. (1995), 'Content validity in psychological assessment: A functional approach to concepts and methods', Psychological Assessment 7(3), 238--247.

Jacoby, J. & Kaplan, L. B. (1972), „The Components Of Perceived Risk", in M. Venkatesan, ed., Proceedings, Third Annual Conference, College Park, ED, Association for Consumer Research, 382-393.

Lawshe, C. (1975), „A quantative approach to content validity", Personnel Psychology 28(4), 563-575.

Lynn, M. (1986), 'Determination and quantification of content validity.', Nursing research.

Mitchell, V.-W. (1999), 'Consumer Perceived Risk: Conceptualisations And Models', European Journal of Marketing 33, 163-195(33).

Nunnally, J. (1967), Psychometric theory, Tata McGraw-Hill.

Storey, V., Straub, D., Stewart, K. & Welke, R. (2000), „A conceptual investigation of the e-commerce industry", Communications of the ACM 43(7), 117-123.

Straub, D.; Boudreau, M. & Gefen, D. (2004), 'Validation guidelines for IS positivist research', Communications of the Association for Information Systems 13(24), 380--427.

Waltz, C. & Bausell, R. (1981), Nursing research: Design, statistics, and computer analysis, FA Davis Company.

Waltz, C.; Strickland, O. & Lenz, E. (2010), Measurement in nursing and health research, Springer Publishing Company.

This paper is supported by the Sectoral Operational Programme Human Resources Development (SOP HRD), financed from the European Social Fund and by the Romanian Government under the contract number POSDRU/89/1.5/S/56815.