Detecting Biased Items When Developing a Scale: A Quantitative Method

Jean-Charles Pillet, Claudio Vitari, Federico Pigni, Kevin Carillo

Detecting Biased Items When Developing a Scale: A Quantitative Method

Completed Research

Jean-Charles Pillet
Grenoble Ecole de Management
Université Savoie Mont Blanc - IREGE
jean-charles.pillet@grenoble-em.com

Claudio Vitari
IAE Paris 1 Panthéon-Sorbonne
(Sorbonne Business School)
vitari.iae@univ-paris1.fr

Federico Pigni
Grenoble Ecole de Management
Univ. Grenoble Alpes ComUE
federico.pigni@grenoble-em.com

Kevin Carillo
Toulouse Business School
k.carillo@tbs-education.fr

Abstract

In survey research, it is well known that the quality of responses is significantly altered by apparently trivial variations in the linguistic or grammatical properties of survey items. Yet numerous seemingly minor changes are made to survey items in the course of the scale development process so that they comply with other requirements (e.g., content validity). As a result, researchers may inadvertently introduce systematic measurement error that is not accounted for in the final model. Remedies to this problem are widely known, but reliable methods to diagnose it do not readily exist. In an effort to address this shortcoming, we develop a quantitative method to detect biased items and reinforce the reliability of IS measurement instruments. In this paper, we provide step by step implementation guidelines and show how to apply the method and interpret the output results.

Keywords

Quantitative method, scale development, item bias, common method bias, ANOVA.

Introduction

Common method bias (CMB) is an ongoing concern among IS researchers and in the social and behavioral science in general (P. M. Podsakoff et al., 2003). CMB comprises all sources of systematic measurement errors that threaten the validity of a study's findings and can lead to common method variance that yields potentially misleading conclusions, because it inflates or deflates relationships among constructs (Campbell & Fiske, 1959; Fuller et al., 2016). By reviewing researches published in top IS journals, Woszczynski & Whitman (2004) have determined that the validity of more than 50% of these studies is jeopardized by common method variance. Recent research efforts have demonstrated empirically that common method variance represents a major validity threat in IS research (Sharma et al., 2009).

Item bias is one of the four sources of common method bias, together with common rater effects, item context effects and measurement context effects that can lead to common method variance (Podsakoff et al., 2003). Indeed, the specific properties or characteristics that an item possesses can alter the cognitive processes involved in survey responding (Tourangeau et al., 2000). For example, respondents who are faced with ambiguous items (e.g., double-barreled, equivocal or unclear items) are likely to engage in random

responding or to resort on their own systematic response tendency (e.g., implicit theories, affectivity, central tendency and leniency biases), which can negatively affect the quality of survey data (Fowler, 1992). Although the problem of item bias is largely recognized, there is little guidance as to how it should be addressed. Current item bias assessment practices involve a mix of cognitive interviews, focus groups, or expert judgments but there is little consensus over which approach should prevail, and few details about how it should be implemented. We address this methodological gap here.

In this paper, we propose a method designed for the sole purpose of detecting and evaluating the risk of item bias in the early stage of the scale development project. The main rationale is that although post hoc statistical remedies exist, they are difficult to implement and they are not entirely reliable (Conway & Lance, 2010; Richardson & Sturman, 2009). In addition, item bias is a source of common method bias that researchers can potentially fully eliminate (Podsakoff et al., 2003). The method we propose relies on the analysis of variance approach similar to the ANOVA approach to content validation (Hinkin & Tracey, 1999) and can be used synergistically with it.

The paper is structured in two parts. In the first part, we explain the protocol we followed to develop valid measures of item bias. The protocol is made of four steps and led to the generation of 21 different measures. In the second part, we demonstrate the value of the method for researchers by showing when and how it can be applied in the course of their scale development effort. For the purpose of this demonstration, we use the items of a new construct introduced by MIS Quarterly in 2016.

Method Development Protocol

The purpose of our four-step method development protocol is to test the validity of the measures of item bias. In the context of this paper, measurement validity refers to the ability of a given measure to discriminate between biased and unbiased items. We adapt the method proposed by Podsakoff et al. (2012) to this need. Briefly, the validation procedure begins with the development of clear conceptual definitions of the main sources of bias, which are then operationalized into measures. The wording of the measures is available in the Appendix. In parallel, we identify items that we judged as manifesting high and low degree of bias within manuscripts published in leading IS journals. We present raters the randomized list of the identified items, and we ask them to assess the extent to which a given item is biased using the measures. If the measures of item bias are valid, there should be a significant and substantial difference in the rating of the items that have been judged by the authors as biased compared to those that have not. This verification procedure is adapted from a widely adopted and recommended technique for testing the content validity of scale items (e.g., Hinkin & Tracey, 1999; Schriesheim et al., 1993; Yao et al., 2008).

Step 1: Sources of Item Bias and Development of the Measures

Item Ambiguity (AMB)

Item ambiguity refers to the fact that "items that are ambiguous allow respondents to respond to them systematically using their own heuristic or respond to them randomly" (Podsakoff et al., 2003: 882). Item ambiguity is likely the biggest problem in the comprehension stage of the response process (Fowler, 1992; Tourangeau et al., 2000). Respondents faced with ambiguous items develop their own idiosyncratic meaning for them and the response quality is likely to be altered as a result.

Item Social Desirability (SDB)

Item social desirability refers to the fact that "items may be written in such a way as to reflect more socially desirable attitudes, behaviors, or perceptions" (Podsakoff et al., 2003:

882). This property of the items is distinct from the tendency of respondents to behave in a culturally acceptable manner (Crowne & Marlowe, 1960; Paulhus, 2001). The social desirability at the item level is often a natural manifestation of social desirability at the conceptual level. For example, we can expect items of the technology addiction scale (Turel et al., 2011) to elicit biased answers. This implies that item social desirability may only be attenuated (as opposed to eliminated) for some specific constructs.

Item Demand Characteristics (IDEM)

Item demand refers to the fact that "items may convey hidden cues as to how to respond to them" (Podsakoff et al., 2003: 882). These cues may come in the form of emotionally charged terms (Graeff, 2004). The implicit assumptions that a question contains can also influence respondents' recall at the retrieval stage (Tourangeau et al., 2000). These cues may influence respondents who are unsure about their response who would then seek to answer in a way that matches their understanding of the need of the researcher (Schuman & Presser, 1996). Survey fatigue or apathy can also lead respondents to simply agree with the item stem (Krosnick, 1991).

Presumption of Usage (PoU)

These three previous definitions will come with little surprise to the eyes of a seasoned scale developer and consensually acknowledged as source of bias by researchers in social science. The choice of words and structure of an item can introduce a more subtle form of bias that goes beyond the ones listed above. Psychometricians contend that presuppositions, defined as "taken for granted assumptions made by the researcher who assumes that a given condition applies to all respondents", are an important source of method bias (Tourangeau et al., 2000: 41-44). Presuppositions not applying to the respondents alter their ability to efficiently process the information contained in the items (Petty & Cacioppo, 1986), which increases the risk that they adopt an automatic response style such as acquiesce (Watson, 1992). We thus theorize about the existence of an additional source of bias that is specific to the object of research in information systems.

Step 2: Selection of the Items

The objective of this phase is to identify items that exhibit high and low degrees of bias. We chose to use published scale items rather than to manipulate items (i.e., purposively introduce a bias in the wording of a seemingly unbiased item, or rewriting a patently biased item in a non-biased way) on the ground that such manipulations would be difficult to implement. On the other hand, using published scale items allows us to demonstrate that the method is capable of detecting sources of item bias that had gone seemingly unnoticed throughout the scale development stages, thereby reinforcing its practical value. Besides, the large number of scales available allows us to introduce a natural variation in the way that item bias manifests to a greater extent that item manipulation would allow.

Each author explored leading IS journals in search for scales containing potentially biased items. We restricted the search to scales published in the "Basket of Eight" outlets. A preliminary list of 44 items (38 potentially biased and 6 potentially unbiased) from 24 different papers has been constituted by the first author based on the input from the other authors. Attention was paid to include in the list a sufficient number of items that are exemplar of each of the four facets of item bias discussed above. Then each of the four authors independently reviewed the 44 items and flagged those that they deemed as potentially biased on each source of item bias. Each author was blind to the other's rating to neutralize the influence of others on one's judgment. The results were aggregated, and we used the number of times a given item was flagged as biased as an indication of consensus. When at least three of the four authors had flagged the item as biased on one dimension of bias, it was considered for inclusion. Ideal candidates would be items that are

consensually identified as biased on one unique source of item bias to limit the risk of confounding. Non-biased items were selected following the same procedure. In total, 18 items from the initial list of 44 items were selected, and all four dimensions of item bias had three different exemplar items.

Step 3: Pre-test

The preliminary set of measures was pre-tested on four PhD students. One of the four students was familiar with common method bias issues. The pre-test consisted of four sequential exercises corresponding to the four sources of item bias identified in the literature. At the end of each exercise, the participants were prompted to provide an oral feedback about the task they had performed. Following the pre-test, the measures of item bias were reworked in such a way that minimizes unnecessary cognitive effort: (1) to assess the attributes of the items rather than their content, (2) to align the format of response options, and (3) to avoid measures that introduced specific interpretive frames.

Step 4: Test of the Validity of the Measures

Implementation

Participants were recruited via the online crowdsourcing platform "Prolific Academic[1]". The platform was selected because it tends to produce more robust results than other alternatives (Peer et al., 2017), and because authors' previous experiences with conducting cognitively demanding rating tasks had already proven successful. We specified our sample to include only native English individuals aged between 18 and 25, as recommended for this type of task (Hinkin & Tracey, 1999; Schriesheim et al., 1993). We split the task in to two samples due to the cognitive effort required to fulfil the task. The two samples were respectively composed of 61 and 57 respondents. All parameters except the measures of item bias are kept equal across the two samples. Thus, each respondent had to evaluate the exact same items but using instruments that differ between sample 1 and sample 2.

The items to evaluate were presented to the raters in four different randomly appearing blocks, to prevent response-order effects (Krosnick & Alwin, 1987). Contextual information was provided at the beginning of each block. Each block included three items identified as biased and two non-biased items, and here as well the order of appearance has been randomized. Several attention questions have been spread throughout the test. Self-reported responses collected at the end of the task indicated that the task was perceived as moderately difficult, and about half the respondents found the task either "somewhat difficult" or "extremely difficult". This was correlated with the written feedback that participants were prompted to provide at the end of the test. One third of the participants reported being either not confident at all or slightly confident in the response they provided in the test. Eliminating these responses did not have any significant influence on the results and were included in the final sample. However, five raters in total failed to spot the attention questions, leaving us with two viable samples of 59 and 57 raters for sample 1 and 2 respectively.

Analysis

The validity of the item bias measures should result in a significant and substantial difference between the ratings of the items judged as biased and nonbiased. To test whether this is significant, we computed a "biased items average score" variable and an "unbiased items average score" for each of the 24 candidate measures. We then used a one-way repeated ANOVA to compare these variables and determine whether the rating of the biased items differs from the rating of the non-biased items.

[1] https://www.prolific.ac/

The results of the ANOVA are reported in Table 1. Assuming that p. values below .05 provide evidence of validity, we conclude that four measures of Item Ambiguity are valid (M2, M4, M5, M6), all the candidate measures of Social Desirability are valid, all measures of Item Demand except M1 are valid, and all measures of Presumption of Usage source of bias are valid. In total, 21 of the 24 tentative measures that have been developed are valid measures of one source of item bias effectively discriminate between biased and unbiased items.

		AMB			**SDB**			**IDEM**			**PoU**		
		Biased	Non-biased	F(df1, df2)	Biased	Non-biased	F(df1, df2)	Biased	Non-biased	F(df1, df2)	Biased	Non-biased	F(df1, df2)
N = 59	M1	2.66	2.70	.03 (ns)	3.50	2.98	8.23**	3.22	3.06	.94(ns)	3.82	3.06	27.61***
	M2	2.70	2.11	14.19***	3.47	2.67	22.33***	3.36	3.01	9.44**	3.71	2.70	67.99***
	M3	2.88	2.75	1.00 (ns)	3.59	2.59	40.12***	3.66	3.27	11.93**	3.34	2.63	21.89***
N = 57	M4	3.11	2.00	63.60***	3.73	2.36	52.63***	3.08	2.66	8.60**	3.16	2.82	7.31**
	M5	2.80	2.13	22.45***	3.82	2.38	68.46***	3.26	2.89	6.36**	3.63	3.06	24.23***
	M6	3.00	1.90	51.67***	3.67	2.38	54.89***	3.51	2.63	41.52***	3.53	2.90	19.23***

Table 1. Test of the validity of the measures of item bias

AMB: Item Ambiguity; SDB: Item Social Desirability; IDEM: Item Demand; PoU: Presumption of Usage.
* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

Application of the Method

We have refined the definition of item bias in the context of IS and developed a set of measures to evaluate the extent to which a given item is biased. We now consider in which situation and how to implement this method in practice. We apply the method to a 9-items scale published in MIS Quarterly in 2016. The scale requires respondents to report their behavior in relation to mobile usage[2]. We choose this scale because items differ greatly in wording (choice of words, grammatical structure, length, etc.), and we suspect that some will perform better than others. The items that we use in the application thus differ from the items that have been mobilized for the specific purpose of validating the measures in the previous section.

Implementation of the Survey

In order to gain a holistic understanding of the scale, we recommend scale developers to evaluate the scale on each source of item bias. Therefore, the first implementation step consists in selecting the measures of item bias. The method development protocol led to the development of four to six different measures for each dimension of item bias, but one measure per source of item bias is probably sufficient in most cases. Indeed, if we accept that the measures of item bias are equally valid measures and that the objective is limited to flagging biased items, there are limited advantages to using multiple measures to assess the same form of bias. For the purpose of this test, we selected the following measures: “It is difficult to intuitively comprehend the statement” for Item Ambiguity; “The statement would elicit a biased answer from respondents who think highly of themselves” for Item Social Desirability; “The statement contains terms that are emotionally loaded” for Item Demand; and “The statement implies that all respondents have used the ICT recently” for Presumption of Usage.

[2] The construct and items are concealed to respect the anonymity of the authors.

The pre-test we conducted indicates that raters would evaluate each dimension of item bias separately rather than simultaneously. Therefore, each block of items to assess should be dedicated to a single measure of item bias. We also recommend randomizing blocks and items, limiting the number of items to evaluate within each block to nine, and to keep the total number of items to evaluate within a reasonable range to prevent cognitive fatigue. Similarly to Hinkin & Tracey (1999), we advise drawing on student as raters for this cognitively demanding task.

Tests and Results

The analysis consists in a systematic comparison between the items' scores and the average score of the scale on each dimension of item bias. In other words, we want to identify items that are rated as significantly more biased than the other items of the scale. To flag biased items, we first compare the average score of the item on a given source of item bias and compare this score to the scale's average, in this case we considered the Grand Mean (GM). When the score of the item is greater than the average score of the scale, the researcher wants to tests whether the difference is statistically significant. To test this, the mean score of the item on a given source of item bias is compared to that of the GM using a one-way repeated ANOVA. Items that score significantly higher than the GM of the scale are susceptible to be biased.

We use IBM SPSS v.23 to run the one-way repeated ANOVA with as many factors as there are items in the scale. In total, four one-way repeated ANOVA are run - one for each measure of item bias. A deviation contrast is conducted so that each item score is compared to the GM. A deviation contrast compares the mean of each level to the mean of all the levels (Grand Mean). The results of these tests are reported in Table 2. For the sake of readability, the cell has been greyed when the item mean score is lower than the scale's GM. The results of the contrast using a Fisher test indicates that item 1 and 5 are significantly more ambiguous than the other items of the scale. Although many items score higher than the scale's average on the social desirability dimension, the difference is not significant enough to conclude that item social desirability is a threat, expect for item 6 (marginally). Item 1, 2, 4 and 6 are perceived by the raters as demanding a certain answer (ie. leading items). Finally, respondents that are not regular users of the ICT may not be able to provide an optimal answer to item 2, 3, 4, and 6.

	AMB		SDB		IDEM		PoU	
	Means	F test	Means	F test	Means	F test	Means	F test
G.M.	**2.66**	N.A.	**3.26**	N.A.	**2.72**	N.A.	**3.38**	N.A.
item1	3.12	F (1,88) = 21.05***	3.23		2.99	F (1,88) = 9.45 **	3.41	F (1,88) = 0.23 (ns)
item2	2.54		3.36	F (1,88) = 2.18 (ns)	2.94	F (1,88) = 7.30 **	3.59	F (1,88) = 13.24 ***
item3	2.56		3.15		2.64		3.58	F (1,88) = 7.60 **
item4	2.58		3.22		2.97	F (1,88) = 8.79 **	3.58	F (1,88) = 11.29 ***
item5	3.32	F (1,88) = 33.43***	2.99		2.63		3.27	
item6	2.60		3.40	F (1,88) = 3.31 +	2.93	F (1,88) = 5.47 **	3.58	F (1,88) = 12.68 ***
item7	2.40		3.33	F (1,88) = 0.66 (ns)	2.49		3.27	
item8	2.60		3.34	F (1,88) = 0.73 (ns)	2.58		3.05	
item9	2.26		3.30	F (1,88) = 0.19 (ns)	2.27		3.05	

Table 2. Comparison of the Mean Item Scores with the scale's Grand Mean (GM)

AMB: Item Ambiguity; SDB: Item Social Desirability; IDEM: Item Demand; PoU: Presumption of Usage.
$^{+}$ $p < 0.10$; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

Summary

Based on these tests, summarized in Table 4, we conclude that there is room for improving the wording of some of the items. Specifically, item 6 has been flagged on three of the four source of item bias and should be specifically scrutinized. On the other hand, items 7, 8 and 9 appear to be free of the main sources of bias identified in the literature.

	AMB	*SDB*	*IDEM*	*PoU*	*Total*
item1	x		x		2
item2			x	x	2
item3				x	1
item4			x	x	2
item5	x				1
item6		x	x	x	3
item7					0
item8					0
item9					0

Table 4. Summary of the test

AMB: Item Ambiguity; SDB: Item Social Desirability;
IDEM: Item Demand; PoU: Presumption of Usage.

Discussion

In this paper, we reviewed the literature on item bias and identify four sources of method bias that can be addressed by improving item wording: item ambiguity, items social desirability, item demand characteristics, and presumption of usage. The first three have been extensively discussed in the literature, but presumption of usage is novel concept that we introduce here. We explain the protocol that has been followed to develop the measures of these four sources of item bias. In total, we developed a total of 21 valid measures. Finally, we demonstrate the value of the method by applying it to a scale published in a leading IS journal. We show that in implementing our tool, researchers will be able to reliably detect underperforming items.

Several limitations of the method have to be outlined. First, it addresses only a limited portion of method bias, namely item characteristic effects. Podsakoff et al. (2003) identified other sources of method bias that are not addressed here, such as common rater effects (e.g., acquiescence bias, consistency bias, mood state, etc.), item context effects (e.g., scale length), and measurement context effects (e.g., measuring predictor and criterion variables at the same point in time). These sources of bias are beyond the scope of the paper yet may account for a significant amount of common method variance (Campbell & Fiske, 1959). A second critic of the method is that it does not capture all the potential sources of item bias. For example, items may carry presuppositions that can be a source of confusion, but we do not capture this aspect specifically. Similarly, equivocal statements that have multiple meanings might be considered appropriate by the raters who perceive a unique meaning, and thus have no problem processing them. Due to these limitations, we urge scale developers to use this method as a complement rather than as a substitute to expert judgment. We believe that the combination of the qualitative insight from experts and the quantitative input from this method can help improve the validity of our measures.

Conclusion

The method we propose fills a methodological gap in the early stages of scale development, during which reworking the items is not as costly as in later stages. It can help scale development scholars to (1) identify the underperforming items of a scale in order to reinforce its psychometric properties, and (2) evaluate the overall risk of method variance that is caused by the properties of the items of the scale. The method consists in comparing the score of each of the items of the scale with one another and with baseline items using one-way repeated ANOVA tests. It has minimal requirements since the rating of the items can be conveniently collected (Peer et al., 2017), and all the tests can be performed using free or common statistical packages. We recommend using the method in conjunction or right after the content validity assessment stage of scale development. In that regard, it could be combined with the method developed by Hinkin and colleagues as respondents are required to rate the measures of a scale individually (Hinkin, 1995; Hinkin & Tracey, 1999). While by no means we imply that this method is a substitute to qualitative techniques (e.g., cognitive interviews, focus groups, card sorting, etc.) that offer more room for creativity and thoughts, it represent a new tool scale developer may leverage to reinforce the validity of future measurement instruments.

REFERENCES

Bhattacherjee, A. (2002). Individual Trust in Online Firms: Scale Development and Initial Test. *Journal of Management Information Systems*, *19*(1), 211–241.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81–105.

Conway, J. M., & Lance, C. E. (2010). What reviewers should expect from authors regarding common method bias in organizational research. *Journal of Business and Psychology*, *25*(3), 325–334.

Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology*, *24*(4), 349–354.

Doty, D. H., & Glick, W. H. (1998). Common methods bias: does common methods variance really bias results? *Organizational Research Methods*, *1*(4), 374–406.

Fowler, F. J. (1992). How unclear terms affect survey data. *Public Opinion Quarterly*, *56*(2), 218–231.

Fuller, C. M., Simmering, M. J., Atinc, G., Atinc, Y., & Babin, B. J. (2016). Common methods variance detection in business research. *Journal of Business Research*, *69*(8), 3192–3198.

Goldberg, L. (1963). A Model of Item Ambiguity in Personality Assessment. *Educational and Psychological Measurement*, *23*(3), 467–492.

Graeff, T. R. (2004). Response Bias. In *Encyclopedia of Social Measurement (Vol. 1)* (pp. 411–418).

Hinkin, T. R. (1995). A Review of Scale Development Practices in the Study of Organizations. *Journal of Management*, *21*(5), 967–988.

Hinkin, T. R., & Tracey, J. B. (1999). An Analysis of Variance Approach to Content Validation. *Organizational Research Methods*, *2*(2), 175–186.

Jackson, D. N. (1964). Desirability Judgments as a Method of Personality Assessment. *Educational and Psychological Measurement*, *24*(2), 223–238.

Johnson, J. A. (1986). Ambiguity, subtlety, and validity of items. In *57th Annual Meeting of the Eastern Psychological Association*.

Kane, G. C., & Borgatti, S. P. (2010). Centrality-Is Proficiency Alignment and Workgroup Performance. *MIS Quarterly*, *35*(4), 1063–1078.

Kollmann, T., Häsel, M., & Breugst, N. (2009). Competence of IT Professionals in E-Business Venture Teams: The Effect of Experience and Expertise on Preference Structure. *Journal of Management Information Systems*, *25*(4), 51–80.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of

attitude measures in surveys. *Applied Cognitive Psychology*, *5*(3), 213–236.
Krosnick, J. A., & Alwin, D. F. (1987). An evaluation of a cognitive theory of response-order effects in survey measurement. *Public Opinion Quarterly*, *51*(2), 201.
Paulhus, D. L. (2001). Socially desirable responding: The evolution of a construct. In *The role of constructs in psychological and educational measurement*. Taylor & Francis Group.
Paulhus, D. L., & John, O. P. (1998). Egoistic and moralistic biases in self-perception: The interplay of self-deceptive styles with basic traits and motives. *Journal of Personality*, *66*(6), 1025–1060.
Peer, E., Samat, S., Brandimarte, L., & Acquisti, A. (2017). Beyond the Turk : Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, *70*, 153–163.
Petty, R. E., & Cacioppo, J. T. (1986). The Elaboration Likelihood Model of Persuasion. *Advances in Experimental Social Psychology*, *19*, 123–205.
Podsakoff, N. P., Podsakoff, P. M., MacKenzie, S. B., & Klinger, R. L. (2012). Are We Really Measuring What We Say We're Measuring? Using Video Techniques to Supplement Traditional Construct Validation Procedures. *Journal of Applied Psychology*, *98*(1), 99–113.
Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: a critical review of the literature and recommended remedies. *The Journal of Applied Psychology*, *88*(5), 879–903.
Richardson, H. A., & Sturman, M. C. (2009). A Tale of Three Perspectives : Examining Post Hoc Statistical Techniques for Detection and Corrections of Common Method Variance. *Organizational Methods*, *12*(4), 762–200.
Schriesheim, C., Powers, K., Scandura, T. A., Gardiner, C. C., & Lankau, M. J. (1993). Improving Construct Measurement In Management Research: Comments and A Quantitative Approach for Assessing the Theoretical Content Adequacy of Paper-and-Pencil Survey-Type Instruments. *Journal of Management.*, *19*(2), 385–417.
Schuman, H., & Presser, S. (1996). *Questions and answers in attitude surveys: Experiments on question form, wording, and context*. Sage Publications.
Sharma, R., Yetton, P., & Crawford, J. (2009). Estimating the Effect of Common Method Variance: The Method-Method Pair Technique with an Illustration from TAM Research. *MIS Quarterly*, *33*(3), 473–490.
Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The Psychology of Survey Response. Cambridge University Press*.
Turel, O., Serenko, A., & Giles, P. (2011). Integrating Technology Addiction and Use: an Empirical Investigation of Online Auction Users. *MIS Quarterly*, *35*(4), 1043-A18.
Venkatesh, V., Brown, S., Maruping, L., & Bala, H. (2008). Predicting different conceptualizations of system use: the competing roles of behavioral intention, facilitating conditions, and behavioral expectation. *MIS Quarterly*, *32*(3), 483–502.
Watson, D. (1992). Correcting for acquiescent response bias in the absence of a balanced scale: An application to class consciousness. *Sociological Methods & Research*, *21*(1), 52–88.
Woszczynski, A. B., & Whitman, M. E. (2004). The Problem of Common Method Variance in IS Research. In A. B. Woszczynski & M. E. Whitman (Ed.), *The Handbook of Information Systems Research* (pp. 66–78). Hershey, PA: Idea Group Publishing.
Yao, G., Wu, C. H., & Yang, C. T. (2008). Examining the content validity of the WHOQOL-BREF from respondents' perspective by quantitative methods. *Social Indicators Research*, *85*(3), 483–498.
Zaichkowsky, J. L. (1985). Measuring the Involvement Construct. *Journal of Consumer Research*, *12*(3), 341–352.

Appendix – List of the Measures of Item Bias

X **SUBDIMENSIONS** **MAIN SOURCES** **MEASURES**	**VALIDITY**
Item Ambiguity: items that allow respondents to respond systematically using their own heuristic or respond to them randomly.	
Semantics (Doty & Glick, 1998)	No
The statement measures something that is abstract. Syntax & semantics (Johnson, 1986)	***
The wording of the statement makes it unclear. Semantics (Goldberg, 1963)	No
The range of possible interpretations of the statement is broad. Syntax (Tourangeau et al., 2000)	***
The statement is worded in a complicated manner. Semantics (Tourangeau et al., 2000)	***
The statement has several possible meanings. Syntax & semantics (Tourangeau et al., 2000)	***
It is difficult to intuitively comprehend the statement.	
Item Social Desirability: items written in such a way as to reflect more socially desirable attitudes, behaviors, or perceptions.	
Item attribute (Jackson, 1964)	**
The statement points at an attribute that is generally perceived as desirable or undesirable. Moralistic response (Paulhus & John, 1998)	***
Respondents that are concerned about how others view them are likely to provide a biased answer to the statement. Moralistic response (Paulhus & John, 1998)	***
Respondents who openly agree with the statement would be perceived either positively or negatively by others. Egoistic response (Paulhus & John, 1998)	

***

The statement would elicit a biased answer from respondents who seek to assert their superiority over others.
Egoistic response
(Paulhus & John, 1998)

***

The statement would elicit a biased answer from respondents who think highly of themselves.
Moralistic response
(Paulhus & John, 1998)

***

The statement would elicit a biased answer from respondents who seek recognition from their peers.
Item Demand: items that convey hidden cues as to how to respond to them (i.e., leading item)
Response range
(P. M. Podsakoff et al., 2003)

No

The range of options that respondents are invited to consider when answering the statement is narrow.
Implicit cues
(P. M. Podsakoff et al., 2003)

**

The statement contains implicit clues as to how respondents are expected to answer.
Leadings response
(P. M. Podsakoff et al., 2003)

**

The personal preference of the person who developed the statement is apparent in its wording.
Leadings response
(Graeff, 2004)

**

The statement is worded in a way that encourages respondents to endorse it.
Implicit cues
(Graeff, 2004)

**

The statement contains words or phrases that imply a certain answer.
Loaded terms
(Graeff, 2004)

***

The statement contains terms that are emotionally loaded.
Presumption of Usage: items that carry implicit assumptions about the respondents that may not apply to all of them.
Interaction with the ICT
(Venkatesh et al., 2008)

***

The statement implies that all respondents are regular users of the ICT.
Interaction with the ICT
(Venkatesh et al., 2008)

***

The statement implies that all respondents have used the ICT recently.
Respondents attributes
(Kane & Borgatti, 2010)

***

The statement implies that all respondents are proficient with the ICT. Respondents attributes (Kollmann et al., 2009)	**
The statement implies that all respondents are knowledgeable about the ICT. Respondents attributes (Bhattacherjee, 2002)	***
The statement implies that all respondents are familiar with the ICT. Interaction with the ICT (Zaichkowsky, 1985)	***
The statement implies that all respondents are personally involved with the ICT.	

Notes: Validity is established based on the performance of the measure and F-statistics (Table 1). No: $p > 0.05$; ** $p < 0.01$; *** $p < 0.001$.