Évaluation de produits technologiques en contexte mobile: premières étapes vers des échelles à items uniques

Nicolas Martin, Eric Jamet, Séverine Erhel, Géraldine Rouxel

Evaluation of Technological Products in Mobility Context: First Steps Toward New Scales with Single Items

Nicolas Martin
b<>com & CRPCC (EA 1285)
1219 Av des Champs Blancs
35510 Cesson-Sévigné, France
nicolas.martin@b-com.com

Éric Jamet
CRPCC (EA 1285) –
Université Rennes 2
Place du Recteur Henri Le Moal
35043 Rennes, France
eric.jamet@univ-rennes2.fr

Séverine Erhel
CRPCC (EA 1285) –
Université Rennes 2
Place du Recteur Henri Le Moal
35043 Rennes, France
severine.erhel@univ-rennes2.fr

Géraldine Rouxel
CRPCC (EA 1285) -
Université Rennes 2
Place du Recteur Henri Le Moal
35043 Rennes, France
geraldine.rouxel@univ-rennes2.fr

http://dx.doi.org/10.1145/3004107.3004133

Résumé
Récolter l'avis d'utilisateurs par questionnaire nécessite généralement l'utilisation de plusieurs items par dimension évaluée pour s'assurer de la fiabilité du questionnaire. Cela peut conduire à de longs questionnaires pouvant décourager les participants. Afin de diminuer cette difficulté, nous présentons ici le développement de deux échelles à items uniques. Ces échelles ont été construites à partir de données issues d'échelles à items multiples testées sur 305 participants évaluant une application mobile. D'après les résultats, l'utilisation d'échelles à items unique ne diminue pas la précision. Cette approche offre de nouvelles possibilités dans l'évaluation, notamment pour les études longitudinales.

Mots Clés
Questionnaire; Échelles à Items Unique; UX; Facteurs Affectivo-Motivationnel.

Abstract
Collecting users' opinions with questionnaires requires generally the use of several items by dimension evaluated to ensure of the tool's reliability. This means long questionnaires which can be discouraging for participants. To deal with this problem, we present here the construction of

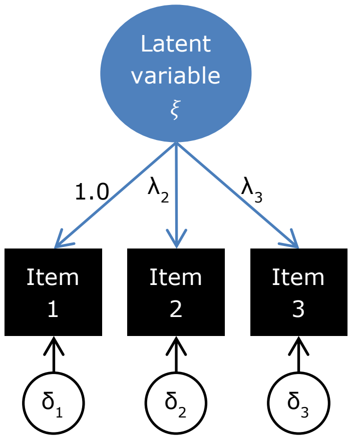

Figure 1. Latent variable structure

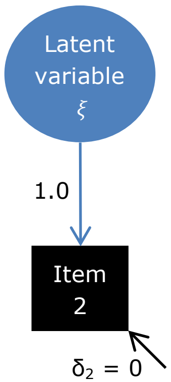

Figure 2. Single item scale: measurement error is fixed to 0 – assumes that item measured the latent variable perfectly

single-items scales. These scales have been constructed from data gathered with multiple-items scales tested on 305 participants who evaluated a mobile app. Results showed no decrease on accuracy compared to a multiple items scales. This approach offers new possibilities for evaluation, especially for longitudinal studies.

Author Keywords

Scale; Single Items; UX; Affective-Motivational Factors.

ACM Classification Keywords

H.5.2. User Interfaces: Evaluation/methodology.

Introduction

Evaluating products with users is a central aim in User-Centered Design approach (UCD) [24]. It requires specific and reliable methods [28] such as user testing, scales, experimental studies, etc. [3]. Thus, scales are often used because they can provide data quickly and they are easy to use. Nevertheless, scales have the disadvantages of requiring several items to ensure of their reliability [13,21]. This can lead to very long scales for participants. Thus, to look after methods to reduce the number of items without decreasing the accuracy seems to be important for the Human Computer Interaction (HCI) field. Indeed, in real-life situations, especially when participants are recruited on social media, their participation is totally based on voluntariness and therefore the length of scales can lead them to give up. In this paper, we will present the application of a method to reduce the length of scales without compromising the precision, through the construction of single-items scales to evaluate factors of acceptability and user experience (UX).

Modeling with latent variables

Generally, in the HCI field, we evaluate theoretical constructs which are difficult to measure directly, for example perceived usability. For this purpose, we typically use scales to assess the observable outcomes of a theoretical variable [13]. From these observable variables, latent variables are often constructed to represent the theoretical variables [8], especially in the Structural Equation Modeling [5,22,27] (SEM) framework. For example, if you want to evaluate perceived usability, several items evaluating this theoretical construct should be developed (e.g., "Using this mobile app seems easy"). Based on data gathered with these items (i.e., observables variables), the latent variable is constructed statistically as the variable which influences the observed covariance between these items, in addition to errors of measure (δ) [5]. However, in order to use latent variables, multiple observable variables (e.g., items) are required. Indeed, the estimation of latent variables and related parameters requires that the model to be identified [25]: this means that the number of unique information (i.e., variances and covariances of observable variables) should be equal or greater than the number of parameters to be estimated (e.g., path between an observable variable and a latent variable or also error of variance of an item) [5]. For example, in the case of a model with one latent variable (as in figure 1) a minimum of three observable variables is required. Indeed, in this case, there are 6 pieces of unique information (variances of the 3 items and covariances between these 3 items) and 6 parameters to be estimated (2 paths between the latent variable and items[1]; measurement errors for the 3 items; variance of the latent variable). Thus, if one item is deleted,

[1] For standardization reasons [5], the variance of the latent variable or the loading of one item needs to be fixed to 1. Here, the loading between the latent variable and the item 1 is fixed.

• Usefulness: I think that this mobile app might be useful

• Usability: The use of this mobile app seems simple

• Stimulation: This mobile app appears innovative

• Trust: I trust the information this mobile app could provide

• Social influence: People that matter to me could encourage me to use this mobile app

• Intrinsic motivation: I used this mobile app because it could be fun to interact with

• Self-Image: I could be positively perceived by others if I used this mobile app

Figure 3. Items selected for the single-items scales

the number of unique information (i.e., variance of the 2 items and covariance between these 2 items = 3 pieces of information) is smaller than the number of parameters to be estimated (i.e., 1 path between the latent variable and the item[1]; errors of variance for the 2 items; variance of the latent variable = 4 parameters to be estimated) and the model becomes not estimable. Thus, the constraint of multiple items can lead to very long questionnaires.

Modeling with Single-Items

Decreasing the number of items so that scales contain only one item per evaluated construct offers practical advantages [7]. Firstly, it can reduce the time spent by participants when testing a new application, and thus facilitates experiments [10]. Moreover, it may reduce the number of missing data due to abandonment of studies. Consequently, it is possible to increase the number of studies, especially longitudinal ones. Indeed, if the participants in a study have to complete only a handful, as opposed to a large number of questions, motivation to participate can increase. Moreover, reducing the number of items provides more flexibility in modeling: the number of parameters to estimate would be then smaller [31]. Nevertheless, as presented previously, statistical models require several items for each evaluated theoretical variable (i.e., latent variable). Despite this requirement, the use of single-items scales (i.e., one item per evaluated theoretical construct) has been developed [e.g. ,1]. Generally, in this case, it can be assumed that the item perfectly measures the theoretical construct [20]. Technically, as presented above, reducing the number of items associated with a latent variable leads to a not estimable model. Therefore, to be able to calculate parameters, some modifications on modeling are carried out as presented in figure 2. First, the loading (λ) between the latent variable and the item is fixed to 1. Secondly, the measurement error (δ) of the item is generally fixed to 0. However, this hypothesis of a perfect measurement is difficult to sustain and can lead to errors in modeling [11]. So, one approach is to fix the measurement error (δ) to a specific non-zero value [5]. With this approach, measurement errors can be incorporated in the estimation of the model to reduce errors in modeling. Technically speaking, when only single-items are used from the creation of the scale, this value is generally chosen from an expected reliability and the variance of item [17]. On the contrary, when the reduction of items is made after a factor analysis on a multiple-items scale, measurement errors of each item can be fixed from estimated measurement errors [20].

Context and current study

As previously mentioned, scales can be good tools to evaluate technological products. Nevertheless, statistical models require several items for each theoretical variable evaluated, meaning that it takes a long time for scales to be completed. To cope with the length of scales, we started the development of short scales including only single items because, to our knowledge, this type of scale can be found in the HCI field. Before going into detail about the creation of scales to evaluate technological products, a short overview of research on factors related to the perception of technological products is presented. Following that, first steps of the development of two short scales with single-items to evaluate technological products are described.

Acceptability and its components

According to previous research, acceptability (i.e., user’s judgement toward a system before use [37]) of a product is based on several factors. Firstly, meta-analyses [e.g. ,26] have shown an influence of usability (i.e., degree in which the user expects the target system to be free of effort [12])

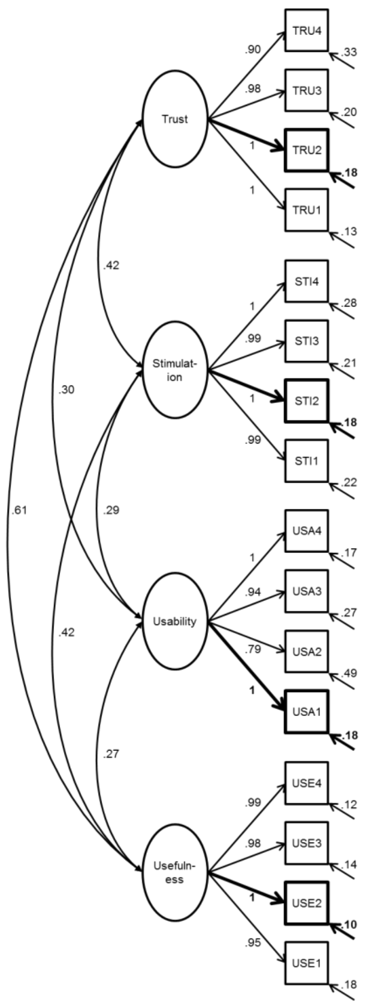

Figure 4. Results of bootstrapped factor analysis for UX scale. Note: Bold line indicate the selected item for single item scale

and usefulness (i.e., degree in which the user expects that using a specific application system will increase his/her performance [12]) on acceptability. More recently, research highlighted that users perceive a product beyond these functional qualities (i.e., beyond usability and usefulness) [4]. Thus, two main variables have been studied: stimulation and trust. Stimulation (i.e., need of novelty [19]) has a positive effect on behavioral intention [36] and the influence of trust (i.e., positive belief about the perceived reliability in a system [16]) has been demonstrated in some papers [e.g. ,18]. These 4 variables are related to product qualities according to user experience (UX) theories. Lastly, in research, the perception toward a technological product is generally considered to be based on a rational evaluation of quality criterions [39]. Nevertheless, the human functioning is more complex and it is necessary to take into account some affective and motivational variables. Thereby, identification (i.e., self-image that the product returns of itself [19]) has been demonstrated as an important predictor of acceptability [36]. Moreover, an effect of intrinsic motivation (i.e., doing an activity for its inherent satisfactions [35]) on acceptability was demonstrated [29]. Lastly, the social influence (i.e., degree in which an individual perceives how important others believe he or she should use the new system [38]) was added and shows a positive effect on behavioral intention [26]. These 3 variables are related to affective-motivational factors. Thus, we developed two short scales to evaluate the theoretical construct presented below: one to evaluate the product qualities (i.e., UX) and one to evaluate affective-motivational factors.

Method

Material

In order to evaluate the psychometrics qualities of our scales, it was necessary to collect data with these scales. Thereby, participants evaluated the sleeping tracker mobile application "Sleep Better ©". This application was selected because it is simple, free, designed for repeated use and relatively distributed.

Measures

To evaluate the users' perceived qualities about the mobile application, the following dimensions were measured:

- <u>Assessment of UX (UX Scale):</u> Usefulness (4 items), Usability (4 items), Stimulation (4 items) and Trust (4 items)
- <u>Assessment of Affective-Motivational Factors (A-M Scale):</u> Intrinsic Motivation (5 items), Social Influence (3 items) and Self-image (4 items)
- Intention to use (4 items)

Some items were constructed specifically for this study and others were based on existing questionnaires as UTAUT [38] and Attrackdiff [19]. For each item, participants had to answer on an 11 points Likert scale (from 0 to 10). Position of each item in the questionnaire was randomly delivered to avoid effects related to order of the items [6].

Participants

Using a web form, 305 participants (236 women), who had never used the app, completed evaluations before use (i.e., 32 items). They were recruited using social networks: Facebook, Twitter and LinkedIn. The average age was 28.83 years (SD = 9.35).

Procedure

The participants completed the scales after reading the presentation of the application proposed by Runtastic© on market app (AppStore and Google Play).

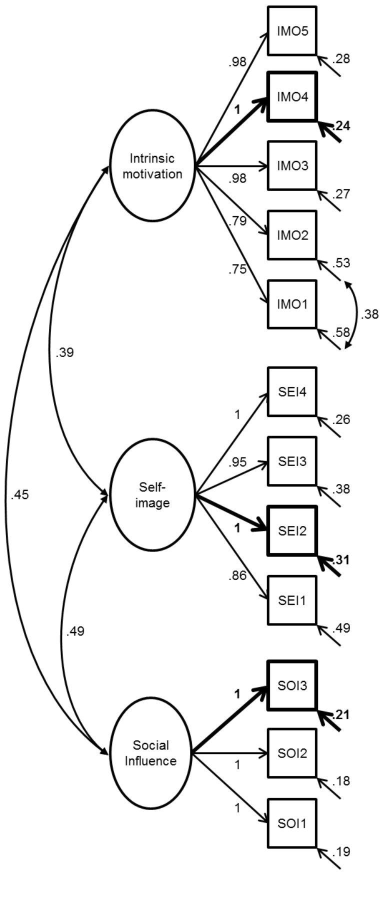

Figure 5. Results of bootstrapped factor analysis for AM scale. Note: Bold line indicate the selected item for single item scale

Results

To ensure of the quality of the scales, exploratory and confirmatory factor analyses were conducted. R [32] with libraries lavaan [34] and psych [33] were used.

Exploratory factor analyses

To determine the number of factors to extract, Eigenvalues and an oblique rotation (Promax) [2] were used. The number of extracted factors was consistent with the number of predefined factors: 4 for UX scale and 3 for A-M Scale. Exploratory factor analyses confirmed the uni-dimensionality of each dimension in UX and A-M Scales: items saturate above .30 their a priori defined factors (i.e., theoretical factors).

Confirmatory factor analyses

After the exploratory factor analyses, confirmatory factor analyses were computed. To deal with violation of multivariate normality, robust standard errors (Huber-White) and a scaled test statistic were used (Yuan-Bentler) [5]. Moreover, for handling with missing values, a full information maximum likelihood estimation was used [14]. According to cutoff values by Hu and Bentler [23], results indicated a good fit for UX and A-M scales with respectively $\frac{\chi^2}{df}$ = 1.36 & 1.64; CFI = .99 & .98; RMSEA = .03 & .05 and SRMR = .03 & .04

Construction of single-items scales

As explained previously, to reduce the length of scales without decreasing their accuracy, one approach is to take into account the measurement errors of items. For this purpose, values of measurement errors should be determined. Thus, confirmatory factor analyses were bootstrapped (10000 draws) to ensure of the stability of models' parameters [15]. This method allows data to be resampled (random sampling with replacement) [9]. Moreover, the observable variables (i.e., items) were standardized before being entered in the model. To develop short versions of the scales, we started by selecting one item per evaluated latent variable. This selection was based on statistical indexes (largest factor loadings and smaller estimated measurement errors) and wording of the items. Selected items are shown in figure 3. Then the estimated measurement errors for these selected items were extracted from bootstrapped factor analyses. Estimated values are presented on figure 4 and 5. With these values, it is possible to correct data collected with items, taking into account the errors of measure.

Multiple-items scales vs. single-items scales

To test the accuracy of single-items scales compared to multiple-items scales, we computed two models which explained Behavioral Intention (BI) using SEM. In both of these models, we estimated the influence of UX and A-M factors on BI. In the first model, scales composed of multiple items for each evaluated construct were used: 16 items for UX scale and 12 items for A-M scale. In the second model, scales composed of single items for each evaluated construct were used: 4 items for UX scale and 3 items for A-M scale. In this model, estimated measurement errors extracted from bootstrapped confirmatory factorial analyses were included to take into account the measurement error. Lastly, to be able to compare the models, the latent variable corresponding to BI was estimated using 4 items. Results indicated good and similar fit with respectively $\frac{\chi^2}{df}$ = 1.49 & 1.6; CFI = .97 & .99; RMSEA = .04 & .05 and SRMR = .04 & .01. Moreover, the paths between BI and UX or Affective-Motivational factors were very similar between scales as presented in figure 6 and 7. Lastly, the explained variances with multiple-items and

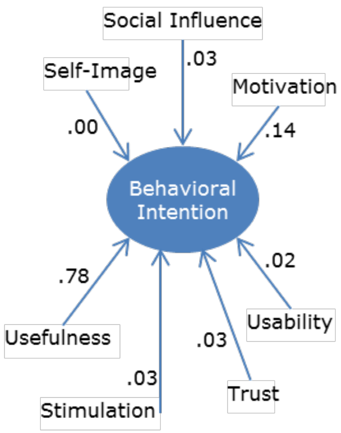

Figure 6. Modeling with multiple-items scales

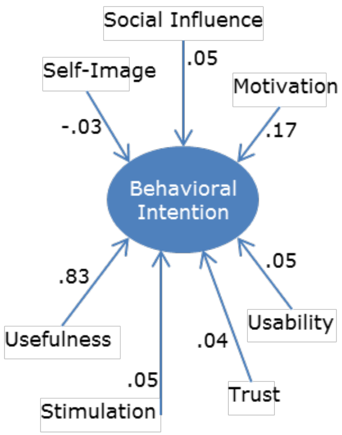

Figure 7. Modeling with single-items scales

single-items scales were similar with respectively 88.7% and 91.6%.

Discussion and conclusion

In the current study, we presented the first steps of the development of short scales including single items. Before creating our short scales, a first version with multiple items for each evaluated dimension was developed. After, one item for each dimension was selected and measurement errors for each item were defined from bootstrapped confirmatory factor analyses. According to the results of SEM, the model which explained BI using single-items scales provides very close results to a model using multiple-items scales. Indeed, reducing the number of items (28 vs. 7 items) has a very little effect on paths between factors of acceptability and behavioral intention. Moreover, the explained variances are close between the both models.

Limits and future studies

Several limits and improvements can be identified. Firstly, the approach presented to take into account the error of measurement has a major disadvantage. Indeed, before being able to use only single items, it is necessary to collect data from participants with multiple items scales to estimate measurement errors. Moreover, as measurement errors are estimated on only one sample, estimated parameters can be biased. To assess this potential bias, future works should strive to collect data with multiple-items scales on several samples or technological products, as well as determining the difference between estimated errors on the first sample and new data. In addition, these data can be used to estimate an average error measurement for each item. Secondly, actual versions of scales include only items for a judgement before use. In order to have a complete tool to evaluate a technological product, it is essential that it can be evaluated after use. Lastly, as scales have been tested in French, it is necessary for use in an international context to translate it into English and revalidate it with English-speaking people.

To conclude this preliminary work, reducing the number of items should decrease the time of completion for participants without decreasing the possibility for statistical modeling and the accuracy of models. We can obtain data from users about a product more easily and then facilitate a user-centered design approach. Moreover, understanding factors leading to adoption and long term use is essential for companies who develop products. For this purpose, it is necessary to acquire data at several stages: before use, just after first use, and over use (e.g. after one week, one month, etc.) [30]. This requires participants to complete scales regularly. Thus, having scales with a minimum of items can facilitate the collection of data for researchers and simplify responses for participants.

Acknowledgements

We would like to thank all those who participated in any way in this research. This work was carried out within the IRT b<>com (ANR-07-A0-AIRT).

References

1. Ahmed M. Abdel-Khalek. 2006. Measuring Happiness with a single-item scale. *Social Behavior and Personality: an international journal* 34, 2: 139–150. https://doi.org/10.2224/sbp.2006.34.2.139
2. Hervé Abdi. 2003. Factor Rotations in Factor Analyses. In *Encyclopedia of Social Science Research Methods*, Michael Lewis-Beck, Alan Bryman and Tim Futing Liao (eds.). SAGE Publications Ltd, Thousand Oaks, Calif.

3. Chadia Abras, Diane Maloney-Krichmar, and Jenny Preece. 2004. User-centered design. *Bainbridge, W. Encyclopedia of Human-Computer Interaction. Thousand Oaks: Sage Publications* 37, 4: 445–456.
4. Javier A. Bargas-Avila and Kasper Hornbaek. 2011. Old Wine in New Bottles or Novel Challenges? A Critical Analysis of Empirical Studies of User Experience. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '11), 2689–2698. https://doi.org/10.1145/1978942.1979336
5. A. Alexander Beaujean. 2014. *Latent Variable Modeling Using R: A Step-by-Step Guide*. Routledge, New York.
6. J. Edwin Benton and John L. Daly. 1991. A Question Order Effect in a Local Government Survey. *Public Opinion Quarterly* 55, 4: 640–642.
7. Lars Bergkvist and John R Rossiter. 2007. The Predictive Validity of Multiple-Item Versus Single-Item Measures of the Same Constructs. *Journal of Marketing Research* 44, 2: 175–184. https://doi.org/10.1509/jmkr.44.2.175
8. Kenneth A Bollen. 2002. Latent variables in psychology and the social sciences. *Annual review of psychology* 53, 1: 605–634.
9. Paula Fitzgerald Bone, Subhash Sharma, and Terence A. Shimp. 1989. A Bootstrap Procedure for Evaluating Goodness-of-Fit Indices of Structural Equation and Confirmatory Factor Models. *Journal of Marketing Research* 26, 1: 105. https://doi.org/10.2307/3172673
10. Timo Christophersen and Udo Konradt. 2011. Reliability, validity, and sensitivity of a single-item measure of online store usability. *International Journal of Human-Computer Studies* 69, 4: 269–280. https://doi.org/10.1016/j.ijhcs.2010.10.005
11. David A. Cole and Kristopher J. Preacher. 2014. Manifest variable path analysis: Potentially serious and misleading consequences due to uncorrected measurement error. *Psychological Methods* 19, 2: 300–315. https://doi.org/10.1037/a0033805
12. Fred D. Davis. 1989. Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Quarterly* 13, 3: 319–340.
13. Robert F. DeVellis. 2017. *Scale development: theory and applications*. SAGE, Los Angeles.
14. Craig K. Enders and Deborah L. Bandalos. 2001. The Relative Performance of Full Information Maximum Likelihood Estimation for Missing Data in Structural Equation Models. *Structural Equation Modeling: A Multidisciplinary Journal* 8, 3: 430–457. https://doi.org/10.1207/S15328007SEM0803_5
15. X. Fan. 2003. Using Commonly Available Software For Bootstrapping In Both Substantive And Measurement Analyses. *Educational and Psychological Measurement* 63, 1: 24–50. https://doi.org/10.1177/0013164402239315
16. B. J. Fogg and Hsiang Tseng. 1999. The Elements of Computer Credibility. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '99), 80–87. https://doi.org/10.1145/302979.303001
17. Christoph Fuchs and Adamantios Diamantopoulos. 2009. Using single-item measures for construct measurement in management research. *Die Betriebswirtschaft* 69, 2: 195.
18. David Gefen, Elena Karahanna, and Detmar W. Straub. 2003. Trust and TAM in Online Shopping: An Integrated Model. *MIS Quarterly* 27, 1: 51–90.
19. Marc Hassenzahl. 2004. The Interplay of Beauty, Goodness, and Usability in Interactive Products. *Human-Computer Interaction* 19, 4: 319–349. https://doi.org/10.1207/s15327051hci1904_2

20. Leslie A. Hayduk and Levente Littvay. 2012. Should researchers use single indicators, best indicators, or multiple indicators in structural equation models? *BMC Medical Research Methodology* 12, 1: 1–17. https://doi.org/10.1186/1471-2288-12-159
21. Timothy R. Hinkin, J. Bruce Tracey, and Cathy A. Enz. 1997. Scale Construction: Developing Reliable and Valid Measurement Instruments. *Journal of Hospitality & Tourism Research* 21, 1: 100–120. https://doi.org/10.1177/109634809702100108
22. Rick H. Hoyle. 2012. *Handbook of Structural Equation Modeling*. Guilford Press, New York.
23. Li-tze Hu and Peter M. Bentler. 1999. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal* 6, 1: 1–55. https://doi.org/10.1080/10705519909540118
24. ISO. 2010. *ISO 9241-210:2010 - Ergonomics of human-system interaction – Part 210: Human-centred design for interactive systems*.
25. D A Kenny and Stephanie Milan. 2012. Identification: A non-technical discussion of a technical issue. In *Handbook of Structural Equation Modeling* (1st ed.), Rick H. Hoyle (ed.). Guilford Press, New York, 145–163.
26. William R. King and Jun He. 2006. A meta-analysis of the technology acceptance model. *Information & Management* 43, 6: 740–755. https://doi.org/10.1016/j.im.2006.05.003
27. Rex B. Kline. 2010. *Principles and Practice of Structural Equation Modeling, Third Edition*. The Guilford Press, New York.
28. Effie L.-C. Law and Paul van Schaik. 2010. Modelling user experience – An agenda for research and practice. *Interacting with Computers* 22, 5: 313–322. https://doi.org/10.1016/j.intcom.2010.04.006
29. Matthew K. O. Lee, Christy M. K. Cheung, and Zhaohui Chen. 2005. Acceptance of Internet-based learning medium: the role of extrinsic and intrinsic motivation. *Information & Management* 42, 8: 1095–1104. https://doi.org/10.1016/j.im.2003.10.007
30. Nicolas Martin, Eric Jamet, Séverine Erhel, and Géraldine Rouxel. In press. From Acceptability to Acceptance: Does Experience with the Product Influence User Initial Representations? In *HCI International 2016 - Posters' Extended Abstracts*. Springer International Publishing.
31. Daniel L. Oberski and Albert Satorra. 2013. Measurement Error Models With Uncertainty About the Error Variance. *Structural Equation Modeling: A Multidisciplinary Journal* 20, 3: 409–428. https://doi.org/10.1080/10705511.2013.797820
32. R Core Team. 2016. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
33. William Revelle. 2015. *psych: Procedures for Psychological, Psychometric, and Personality Research*. Northwestern University, Evanston, Illinois.
34. Yves Rosseel. 2012. lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software* 48, 2: 1–36.
35. Richard M. Ryan and Edward L. Deci. 2000. Intrinsic and Extrinsic Motivations: Classic Definitions and New Directions. *Contemporary Educational Psychology* 25, 1: 54–67. https://doi.org/10.1006/ceps.1999.1020
36. Paul van Schaik and Jonathan Ling. 2011. An integrated model of interaction experience for information retrieval in a Web-based encyclopaedia. *Interacting with Computers* 23, 1: 18–32. https://doi.org/10.1016/j.intcom.2010.07.002
37. Geertje Schuitema, Linda Steg, and Sonja Forward. 2010. Explaining differences in acceptability before and

acceptance after the implementation of a congestion charge in Stockholm. *Transportation Research Part A: Policy and Practice* 44, 2: 99–109. https://doi.org/10.1016/j.tra.2009.11.005

38. Viswanath Venkatesh, Michael Morris, Gordon Davis, and Fred Davis. 2003. User Acceptance of Information Technology: Toward a Unified View. *Management Information Systems Quarterly* 27, 3.
39. Viswanath Venkatesh and Michael G. Morris. 2000. Why Don't Men Ever Stop to Ask for Directions? Gender, Social Influence, and Their Role in Technology Acceptance and Usage Behavior. *MIS Quarterly* 24, 1: 115–139. https://doi.org/10.2307/3250981