Computers in Human Behavior Reports 1 (2020) 100004

Contents lists available at ScienceDirect

Computers in Human Behavior Reports

journal homepage: www.journals.elsevier.com/computers-in-human-behavior-reports

A multi-method psychometric assessment of the affinity for technology interaction (ATI) scale

Olga Lezhnina *, Gábor Kismihók

Learning and Skill Analytics Research Group, Leibniz Information Centre for Science and Technology University Library, Hannover, Germany

ARTICLE INFO

Keywords:
Human-technology interaction
Affinity for technology interaction
Construct validity
Hierarchical clustering on variables
Psychometric protocol

ABSTRACT

In order to develop valid and reliable instruments, psychometric validation should be conducted as an iterative process that "requires a multi-method assessment" (Schimmack, 2019, p. 4). In this study, a multi-method psychometric approach was applied to a recently developed and validated scale, the Affinity for Technology Interaction (ATI) scale (Franke, Attig, & Wessel, 2018). The dataset ($N = 240$) shared by the authors of the scale (Franke et al., 2018) was used. Construct validity of the ATI was explored by means of hierarchical clustering on variables, and its psychometric properties were analysed in accordance with an extended psychometric protocol (Dima, 2018) by methods of Classical Test Theory (CTT) and Item Response Theory (IRT). The results showed that the ATI is a unidimensional scale (homogeneity $H = 0.55$) with excellent reliability ($\omega = 0.90$ [0.88-0.92]) and construct validity. Suggestions for further improvement of the ATI scale and the psychometric protocol were made.

1. Introduction

In the contemporary world, human interaction with technology is crucially important and has been thoroughly studied in recent decades. Psychometric instruments can be seen as vehicles for replicability and reproducibility of research in this area (Lewis, 2015). These instruments were effectively used in research of users' attitudes toward interactive television (Fulford & Zhang, 1993), virtual reality environments (Huang, Rauch, & Liaw, 2010), teacher attitudes toward computers (Christensen & Knezek, 2009) tablet computers (Pruet, Ang, & Farzin, 2016), or gender differences in attitudes toward computers (Young, 2000).

Unfortunately, many psychometric instruments still have "unclear validity" (Bargas-Avila & Hornbæk, 2011, p.1). Their insufficient validation contributes to "questionable conclusions and difficulty of subsequent research to replicate" (Flake, Pek, & Hehman, 2017, p. 374). More rigorous practices in scale development and validation, as well as more transparent reporting (Hogan & Agnello, 2004), are required to overcome this situation, which Schimmack (2019) in his recent paper called validation crisis.

To resolve this crisis, psychometric analysis should be conducted as an iterative process that "requires a multi-method assessment" (Schimmack, 2019, p. 4). In other words, even when the results of a psychometric study are convincing and exhaustively reported, it would always be beneficial to re-examine the main characteristics of the instrument with more rigorous, or newly developed, or simply different methods. Thus, the two-fold goal is achieved: the scale is more comprehensively validated, and methodology of psychometric research is further developed.

In the current study, a multi-method approach was applied to the ATI scale (Franke et al., 2018), a recently developed and validated scale measuring users' engagement in technology interaction. The authors analysed the data generously shared by Franke et al. (2018) with the intention to re-examine psychometric properties of the scale, on the one hand, and contribute to development of psychometric research methodology, on the other.

2. Theoretical background

2.1. Scale validation practices

In this section, we briefly outlined contemporary validation practices and methodological recommendations relevant to our study. Description of psychometric properties which were not explored in the current study (e.g. criterion validity or test-retest reliability) had to be omitted.

* Corresponding author.
E-mail address: Olga.Lezhnina@tib.eu (O. Lezhnina).

https://doi.org/10.1016/j.chbr.2020.100004
Received 26 November 2019; Received in revised form 30 January 2020; Accepted 30 January 2020
Available online 29 May 2020

2.1.1. Validity

Validity is the crucial characteristic of a scale, without which “all the other measurement characteristics become relatively inconsequential” (Hogan & Agnello, 2004, p. 802). Here we focus on construct validity, in particular, on convergent and discriminant validity. Typically, convergent validity is assessed by correlations between the scores on the scale under study and scores on existing measures for similar constructs, and discriminant validity by correlations between the scores on the scale and the scores on existing measures for conceptually different constructs; bivariate regression analysis is also used for these purposes (Boateng, Neilands, Frongillo, Melgar-Quiñonez, & Young, 2018).

Recently, structural equation modeling (SEM) has been recognized as an effective tool for demonstrating convergent and discriminant validity of an instrument (Schimmack, 2019). However, assumptions for SEM should be met in order for the analysis to be meaningful; it has been repeatedly stressed that ignoring multivariate normality assumption is a major problem in studies using SEM (Goodboy & Kline, 2017).

Hierarchical clustering of variables can be also used as an alternative method of assessing construct validity, as it makes possible to obtain meaningful structures by arranging variables into homogeneous clusters (Chavent, Kuentz, Liquet, & Saracco, 2011). Variables that are strongly related to each other, and thus contain similar information, are united in the same cluster. It can be achieved by agglomerative hierarchical clustering, which starts with each variable forming a separate cluster; the number of clusters is reduced at each stage based on a similarity or dissimilarity criterion, until all units are agglomerated in a single cluster. Çokluk, Büyüköztürk, and Kayri (2010) concluded that agglomerative hierarchical clustering “may be used as an alternative approach in producing additional findings related to the construct validity of scales” (p. 6403).

2.1.2. Item analysis

In addition to exploring validity of a scale, the researcher needs to study its item-level functioning. In the frame of Classical Test Theory (CTT), item-level descriptives and scores distributions are explored to assess item difficulties, item discriminations and detect ceiling or floor effects.

Dima (2018) developed an extended psychometric protocol, in which item properties are assessed both with CTT and Item Response Theory (IRT), a nonlinear probabilistic model that is less simplified than CTT. In Table 1 in Dima’s paper (Dima, 2018, p. 145), the main procedures of the psychometric protocol are summarised, theoretical background briefly explained and decision criteria given.

Item properties in the frame of IRT are characterised by item fit and person fit measures (Wang et al., 2017), infit and outfit values, Item Characteristic Curves, etc. It was emphasized that IRT should be used more widely to explore psychometric properties of scales, as it is “consistent with a cognitive theory of how people respond to questions” (Singh, 2004, p. 205). Moreover, IRT allows maintaining the width of the latent continuum and diagnosing whether the test is able to differentiate between the respondents’ ability on the latent dimension, which decreases the occurrence of Type II error (Dima, 2018).

2.1.3. Dimensionality

Exploratory factor analysis (EFA) is the most common tool for studying the dimensionality of a scale (Field, Miles, & Field, 2012); it has, however, potential pitfalls that need to be avoided. Howard (2016) gave guidelines on best practices for EFA. In addition to sample size considerations and recommendations on factor loading cut-offs, they include checking assumptions for EFA with Barlett’s test of sphericity and the Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy; choosing Principal Axis Factoring (PAF) or Maximal Likelihood (ML) rather than Principal Component Analysis (PCA) as a factor analytic method; and applying scree plot analysis for factor retention in combination with either parallel analysis, or Velicer’s Minimum Average Partial (MAP) test, or both, rather than Kaiser criterion.

When the structure of a scale is supposed to be known from previous studies, it is usually checked with confirmatory factor analysis (CFA). Neglecting check of assumptions for the CFA, such as multivariate normality assumption (Goodboy & Kline, 2017), and insufficient information on model fit at the stage of reporting (Jackson, Gillaspy, & Purc-Stephenson, 2009) are well known problems of the CFA studies.

Dimensionality of a scale can be also explored with Very Simple Structure Analysis (VSS) and Item Cluster Analysis (ICLUST) (Revelle, 1978). VSS assesses the fit of several models of increasing complexity (that is, with the increasing number of factors) using residual matrix of each model, while ICLUST hierarchically clusters items of a scale and visualizes the results as a cluster diagram (Revelle, 2018). Each of these methods can be used in combination with others as an additional source of information.

Mokken Scaling Analysis (MSA) is an effective method of dimensionality assessment, which is still underused in psychometric research (Stochl, Jones, & Croudace, 2012; Watson et al., 2012). MSA relies on nonparametric IRT and includes checking the assumptions of homogeneity, monotonicity, local independence and invariant ordering. In contrast to CFA, it does not imply the restrictive assumption of multivariate normality (Emons, Sijtsma, & Pedersen, 2012) and gives “less biased and more parsimonious estimations of dimensionality compared to the FA” (Dima, 2018, p. 142).

In Dima’s (2018) psychometric protocol, all these methods - MSA, EFA, ICLUST, VSS and CFA - are used to provide comprehensive assessment of dimensionality of a scale. Results of these methods can be compared to reach a conclusion about the structure of the scale.

2.1.4. Reliability

Reliability is one of the most important characteristics of a scale, and all kinds of reliability (test-retest, interrater etc.) give invaluable information to the researcher; here we focus on internal consistency reliability. It is typically estimated with Cronbach’s alpha, which can be reported alongside with the confidence interval (Dunn, Baguley, & Brunsden, 2014). However, McDonald’s omega was shown to be a less biased and more informative estimator than Cronbach’s alpha (Crutzen & Peters, 2017).

In Dima’s (2018) psychometric protocol, Cronbach’s alpha with the confidence interval, McDonald’s omega with the confidence interval, as well as less frequently used reliability indices, Guttman’s lambda-6 and worst split halt reliability (beta), are estimated.

2.2. Affinity for technology interaction scale

In the area of human-technology interaction, even psychometrically sound instruments often ignore individual differences (Schmettow, Noordzij, & Mundt, 2013) or describe users’ interaction with already outdated technology without taking into account rapidly changing digital environment (Attig, Wessel, & Franke, 2017). Therefore, Franke et al. (2018) developed the ATI scale, an economic nine-item instrument that measures affinity for technology interaction (formulations of the items are given in the Appendix). According to the authors of the scale, multiple studies ($n > 1500$) gave “satisfying results with regard to dimensionality, reliability, validity and distribution of ATI score values” (Franke et al., 2018, p. 163).

Affinity for technology interaction was defined by Franke et al. (2018) as “the tendency to actively engage in intensive technology interaction” (p. 456). It is rooted in need for cognition, an established psychological construct, “a stable individual difference in people’s tendency to engage in and enjoy effortful cognitive activity” (Cacioppo, Petty, Feinstein, & Jarvis, 1996, p. 197), which is an influential factor of human information processing: it was shown that high need for cognition is related to use of "strategies such as critical processing, relating and structuring" (Cazan & Indreica, 2014, p. 134).

As a construct, affinity for technology was previously explored by other researchers, but the scales they developed were not sufficiently effective in terms of construct definition or unidimensionality and economy. For instance, Edison and Geissler (2003) defined affinity for technology as "positive affect towards technology" (p. 140), while the scale that they constructed included items related to skills rather than to positive affect, such as "I know how to deal with technological malfunctions or problems" (p. 154). Karrer, Glaser, Clemens, and Bruder (2009) developed a 19-item scale, which includes four factors: "enthusiasm", "competence", "positive consequences", and "negative consequences". Thus, the ATI scale by Franke et al. (2018) has advantages as a unidimensional economic scale measuring a clearly defined construct rooted in an established psychological attribute.

3. Methods

3.1. Dataset

The dataset ($N = 240$) was shared by Franke et al. (2018) and is currently available on their website (Franke, Attig, & Wessel, 2020). The data was collected by means of MTurk in the USA. Demographic variables were not included in the current analysis. There were scores on 12 scales in the dataset: (a) the Affinity for Technology Interaction (ATI, nine items); (b) Technical Problem Solving Success (TPSS, four items); (c) Technical System Learning Success (TSLS, three items); (d) Interest in Technology (Interest, four items); (e) Need for Cognition (NFC, four items); (f) Geekism (GEX, 15 items); (g) a short form of Big Five Inventory (BFI-10, ten items); and (h) Regulatory Focus Scale adapted for technical systems (RFC, six items). There was no missing data in the scales subset due to strict quality filtering, including completeness check. As the authors of the scale reported, those respondents who (a) did not complete their survey, (b) completed the survey twice, (c) failed to answer the built-in attention checks, or (d) resided outside the USA were excluded from the dataset. For more information on the sample, the scales and quality filtering see Franke et al. (2018).

3.2. Data analysis

Data analysis was conducted with R, version 3.5.2 (R Core Team, 2013). A coherent system of packages tidyverse (Wickham, 2017) was used for data manipulation and visualization. The R script is available on GitHub (Lezhnina, 2020).

Construct validity of the ATI scale was assessed by hierarchical cluster analysis. Package ClustOfVar (Chavent et al., 2011) was used for agglomerative hierarchical clustering, the homogeneity criterion of a cluster was defined as the sum of correlations ratios to a synthetic variable (the first component obtained by a principal component analysis), and stability of partitions was evaluated with a bootstrap approach.

According to Dima's (2018) protocol, psychometric analysis was conducted to explore: (a) item descriptive statistics; (b) item properties according to non-parametric IRT, with homogeneity, monotonicity, local independence, and invariant ordering assumptions checked; (c) item properties according to parametric IRT requirements; (d) structure of the scale according to EFA, CFA, VSS and ICLUST; (e) reliability of the scale and item properties according to CTT; (f) score statistics and distributions.

A few amendments were made to Dima's (2018) protocol in the current study (see Discussion). They were related to more explicit checking of assumptions for CFA and EFA. For CFA, multivariate normality of the data was checked with Mardia's test from package QuantPsyc (Fletcher, 2012) and from package MVN (Korkmaz, Goksuluk, & Zararsiz, 2014); for EFA, the KMO and Bartlett's test of sphericity were conducted with psych package (Revelle, 2018), and for factor extraction, nFactors package was used (Raiche, 2010).

4. Results

4.1. Construct validity: hierarchical cluster analysis

Hierarchy of variables was explored for all constructs (means of all scales). Stability of partitions was checked with the default of 100 bootstrap samples. The results, which are presented in Fig. 1, suggested that eight-cluster and nine-cluster partitions were most stable.

The eight-cluster partition was explored: the ATI formed one cluster with Geekism, Interest in Technology and RFS constructs. In case of the nine-cluster partition, the ATI formed one cluster with Geekism and Interest in Technology constructs.

For the minimal possible number of clusters, the two-cluster partition, the results were similar: the ATI was most closely related to such constructs as Geekism and Interest in Technology, with other technology-related constructs being in the same cluster with it, while Big Five dimensions formed another cluster. The gain in cohesion for the two-cluster partition was 20.59%, for the eight-cluster partition 85.96%, and for the nine-cluster partition 91.60%. The dendrogram for the two-cluster partition, with height indicating the values of the aggregation criterion, is presented in Fig. 2.

Hierarchy of variables was also constructed for all items of all scales. The ATI items were close to Geekism, Interest in Technology and Need for Cognition items. Stability of partitions was explored with the default of 100 bootstrap samples. The two-cluster partition showed relatively high stability according to the adjusted Rand indices, with the next stability maximum at 26 clusters. The two-cluster solution was explored: it gave a gain in cohesion of 6.99%, and all items of the ATI scale were close to items of other technology-related scales in one of the clusters, with Big Five items in the other cluster. The 26-cluster partition was explored. The gain in cohesion was 71.19%. Items of the ATI scale formed clusters with items from Geekism, Interest in Technology and Need for Cognition items. Overall, the results of cluster analysis supported the conclusions by Franke et al. (2018) about convergent and discriminant validity of the ATI scale.

4.2. Psychometric protocol (Dima, 2018)

The ATI items showed sufficient variation to differentiate respondents on their affinity for technology interaction. Descriptive statistics of the items is given in Table 1. Skew and kurtosis values were acceptable. Hereinafter, reverse coded items (ati03R, ati06R and ati08R) are indicated with the letter R.

Barplots for items, which show frequencies of endorsement, are presented in Fig. 3. All response options were represented in the data.

Associations between items were positive. Item ati03R showed the weakest correlations with other items, ranging from 0.14 to 0.26, while other items correlated with each other in range from 0.46 to 0.85.

Multivariate normality of the data was studied with Mardia's test as our amendment to Dima's (2018) protocol, as the latter includes check of multivariate outliers based on Mahalanobis D^2 values but not Mardia's test. The results of the test for skew and kurtosis were significant, with $p < .001$, which means that the assumption of multivariate normality was violated.

Aberrant response patterns of respondents were explored with analysis of Guttman errors. There were 13 outliers (cases with a number of Guttman errors higher than the cut off value of 65.5). As there was no valid reason to remove the outliers (Bakker & Wicherts, 2014), the cases were kept in the dataset for further analysis.

In the frame of Mokken Scaling Analysis (MSA), an Automated Item Selection Procedure (AISP) was conducted to explore scalability of items and dimensionality of the scale at increasing threshold levels of homogeneity. According to the AISP results, the ATI scale is unidimensional. As the minimum threshold level for homogeneity is .30, items with value 0 at this level or below are considered unscalable. For the ATI, there was one item (ati03R), which showed lack of scalability at the threshold as

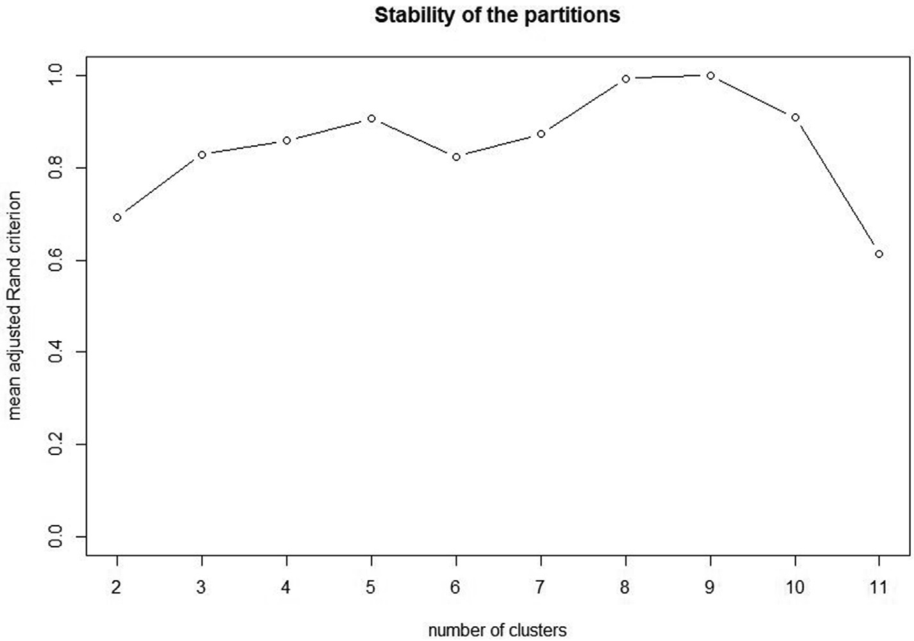

Fig. 1. Stability of partitions for all constructs based on the adjusted Rand indices.

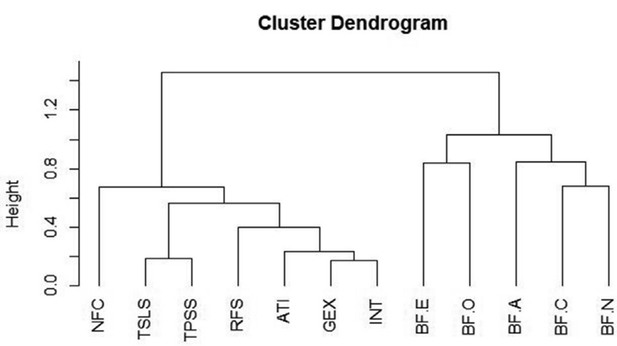

Fig. 2. Dendrogram of all constructs. NFC = Need for Cognition; TSLS = Technical System Learning Success; TPSS = Technical Problem Solving Success; RFS = Regulatory Focus Scale adapted for technical systems; ATI = Affinity for Technology Interaction; GEX = Geekism; INT = Interest in Technology; BF = Big Five scales: BF. E = Extraversion; BF.O = Openness; BF. A = Agreeableness; BF.C = Conscientiousness; and BF.N = Neuroticism.

Table 1
Descriptive statistics of the ATI items.

Items	*M*	*SD*	Skew	Kurtosis	*SE*
ati01	3.91	1.25	−0.50	−0.23	0.08
ati02	4.22	1.24	−0.65	0.02	0.08
ati03R	3.66	1.35	−0.21	−0.81	0.09
ati04	4.27	1.15	−0.57	0.18	0.07
ati05	4.39	1.11	−0.53	0.07	0.07
ati06R	3.50	1.46	0.04	−0.91	0.09
ati07	3.67	1.39	−0.32	−0.83	0.09
ati08R	3.23	1.24	0.26	−0.52	0.08
ati09	4.35	1.14	−0.73	0.51	0.07

low as 0.25. Thus, it could be recommended to remove this item from the scale. Further analysis was conducted for both versions of the ATI scale, the current version (ATI) and the eight-item version with item ati03R removed (hereinafter called ATI8). The results are reported separately whenever the comparison between the two versions is meaningful; in other cases, the results for the original version (ATI) are reported.

The complete item set of the ATI scale had a homogeneity value $H = 0.55$ with a standard error of 0.03, and the complete item set of the ATI8 scale had a homogeneity value $H = 0.64$ with a standard error of 0.03. Thus, removing item ati03R would lead to increase in homogeneity of the whole scale and to increase in homogeneity of all items, as can be seen from Table 2.

According to local independence (conditional associations) test, all nine items of the ATI scale meet the local independence criterion. Monotonicity test (with default minisize of $n = 80$) gave criterion values (Crit) of 0 for all items, except for item ati03R; for this item, the Crit value was 41. As the threshold for the Crit value is 40, and ideally, an item should have the Crit value of 0 (Schwab, Dichter, & Berwig, 2018, p. 4), the monotonicity test for item ati03R showed violation of the assumption, while other items showed very good monotonicity.

Invariant item ordering (IIO) test for the ATI (with the default minisize) showed that there were significant violations of invariant ordering for items ati01, ati06R and ati07 (one violation per each item) and three significant violations of invariant ordering for item ati03R. The output of the test directly suggested removing item ati03R from the scale. When IIO test was conducted for ATI8 scale, it showed zero violations of invariant ordering for each item.

Summary item fit for the ATI scale was explored. Criteria for item fit are the mean squares ranging from 0.6 to 1.4, and values above 2 are considered not suitable for measuring the latent construct on an interval level (Dima, 2018). Outfit and infit mean squares values of all items, except for item ati03R, ranged from 0.56 to 0.95. For item ati03R, the outfit value was 2.52, and the infit value 2.10, which was above the threshold. Thus, all items of the ATI, except for item ati03R, form a scale that satisfies requirements for additive measurement. This result was supported by local independence test (fit on the two ways margins). Person fit was evaluated based on the same criteria as item fit. There were no respondents with misfit according to outfit values or infit values.

The hierarchy of item difficulty and the match between person ability and item difficulty (scale targeting) were explored graphically via Person-Item map (see. e.g., Cappelleri, Jason Lundy, & Hays, 2014). For the ATI scale, separation reliability was 0.91, and person separation was 3.25. With the cut off values of 0.80 for separation reliability and 2 for person separation (Dima, 2018), it means that the ATI is able to differentiate between the respondents regarding their level of affinity for technology interaction. For the ATI8 scale, separation reliability was

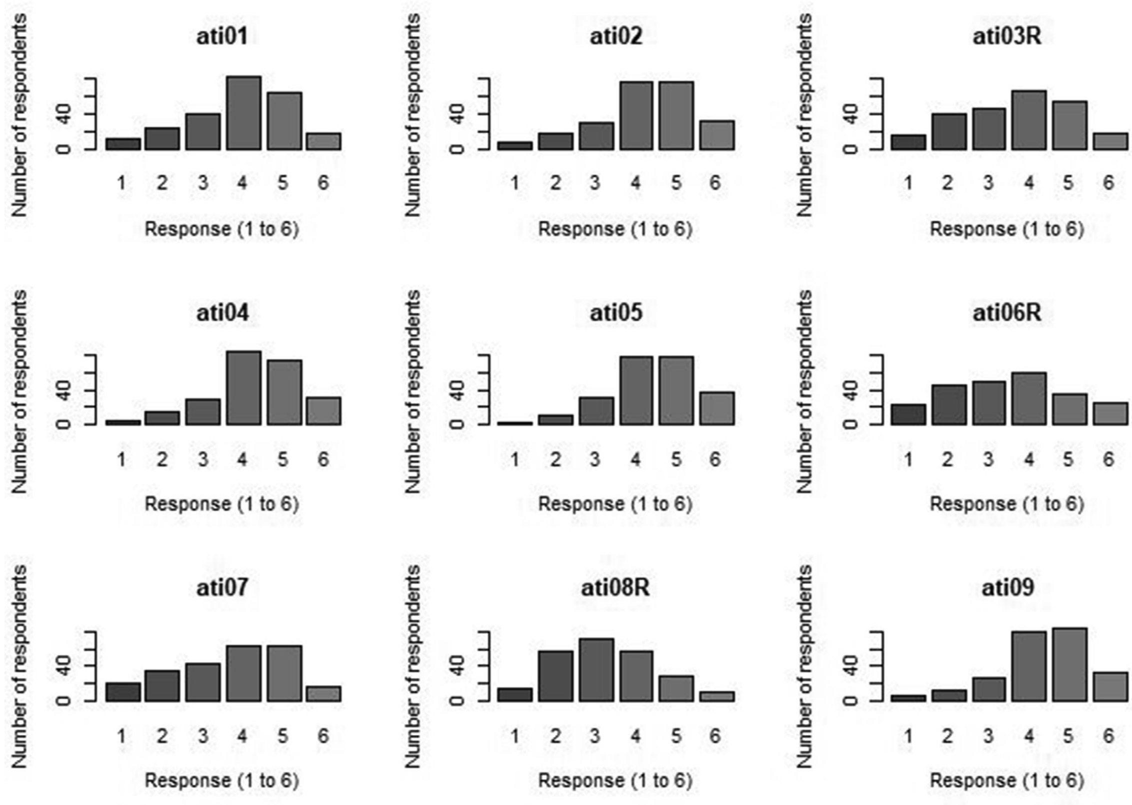

Fig. 3. Barplots for the ATI items.

Table 2
Homogeneity values for items of ATI and ATI8 scales.

Scales	ati01	ati02	ati03R	ati04	ati05	ati06R	ati07	ati08R	ati09
ATI	.57	.61	.23	.58	.63	.57	.60	.54	.60
ATI8	.62	.67	–	.64	.69	.63	.68	.59	.65

0.93, and person separation was 3.73.

In accordance with Dima's (2018) protocol, CFA with robust maximum-likelihood (MLR) estimator was conducted on the ATI scale, with one factor as suggested by the MSA results. The following rules were applied to assess the global fit between the tested model and the data (Hu & Bentler, 1999): the chi square test should reveal no significant differences between the model and the observed covariances ($p \geq .05$), CFI $\geq$ 0.95, RMSEA $\leq$ 0.08 and SRMR $\leq$ 0.08. The results did not show a good fit of the model. The chi square test was significant χ^2 (27) = 231.34, $p <$ .001, CFI = 0.84, RMSE = 0.19, SRMR = 0.07. CFA for the ATI8 scale also showed insufficient fit, and changes based on specification search with modification indices did not significantly improve the fit. It can be explained by the fact that the data was not multivariate normal, and CFA, even with a robust estimator, was not recommendable in this case.

An exploratory factor analysis (EFA) was conducted on the ATI scale. The KMO verified the sampling adequacy for the analysis (for the scale, KMO = .89), with all KMO values for individual items above 0.85. Bartlett's test of sphericity, χ^2 (36) = 1448.93, was significant with $p < .001$, thus assumptions for the EFA were met. Principal Axis Factoring (PAF) was used as a factor analytic method. To determine the number of factors, parallel analysis, scree plot analysis and Velicer's Minimum Average Partial (MAP) test were used. The results are presented in Fig. 4. It can be seen that all methods support one-factor solution (acceleration factor gives a numeric expression to scree plot inspection results).

The decision to retain one factor was supported by VSS analysis, which indicated that the first level of complexity achieves a maximum of 0.92 with one factor. The Velicer's MAP achieved a minimum of 0.06 with one factor. ICLUST also gave one cluster solution. Therefore, one factor, which explained 55% of variance, was retained in the final analysis. Standardized factor loadings for all items except one ranged from 0.68 to 0.85, while for item ati03R the loading was 0.28.

Results of reliability analysis the ATI and ATI8 scales are presented in Table 3. They include Cronbach's alphas with confidence intervals, McDonald's omegas with confidence intervals, Guttman's lambdas-6 and worst split halt reliabilities (betas). Both scales, the original ATI and the ATI8, had excellent reliability according to all indices.

Results of item analysis of the ATI scale are reported in Table 4. Item discriminations (corrected item-total correlations) and Cronbach's alphas when the item is removed are presented for all items of the scale.

5. Discussion

5.1. Summary of results

The results of our hierarchical cluster analysis supported conclusions by Franke et al. (2018) conceptualizing the affinity for technology interaction as close to geekism and interest in technology constructs and distinct from Big Five dimensions. The results of the psychometric study in accordance with Dima's (2018) psychometric protocol showed that the ATI is a unidimensional scale with good homogeneity (less scalable item ati03R needs to be explored further), good ability to differentiate respondents on the measured construct, satisfying the requirements of

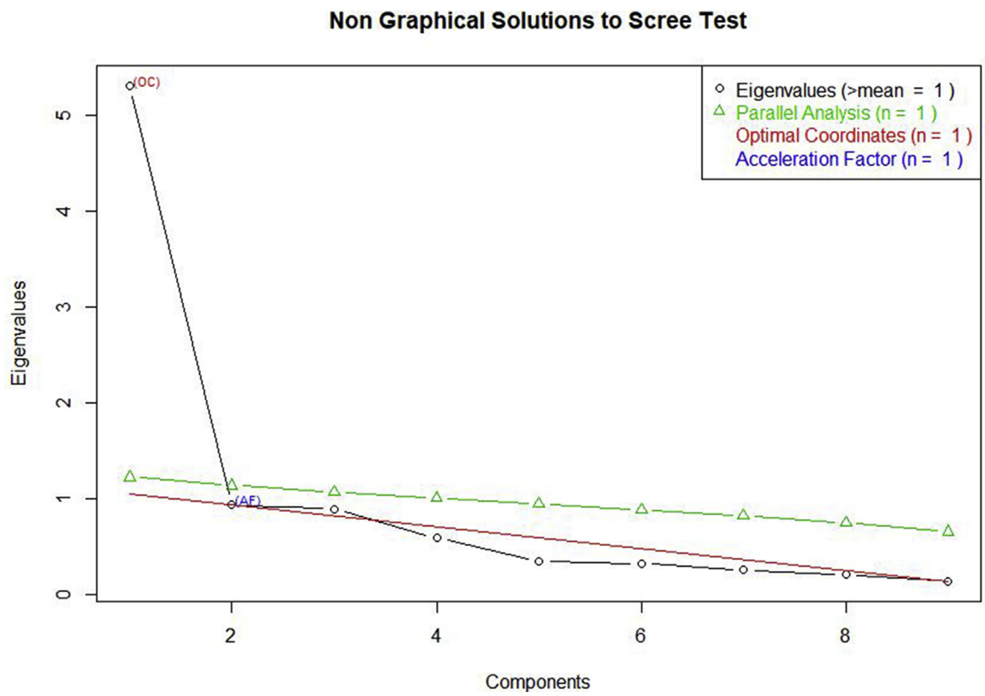

Fig. 4. The scree plot for the ATI scale.

Table 3
Reliability indices for the ATI and the ATI8 scales.

Scale Version	Alpha	Guttman's λ-6	Beta	Omega
ATI	.90 [.88-.92]	.92	.83	.90 [.88-.92]
ATI8	.92 [.91-.94]	.93	.86	.92 [.89-.94]

additive measurement, with good monotonicity and invariant order difficulty at all levels of latent dimension, and with high reliability according to all indices.

5.2. Implications

In the current study, multi-method psychometric procedures were applied to re-analyze validity, dimensionality, and reliability of ATI scale. The implications of the study are thus twofold: First, these are implications related to the ATI scale, and second, to further development of psychometric validation procedures.

As the ATI scale is proved to be a valid and reliable instrument, it could be recommended for use in research of human-technology interaction. This economic unidimensional instrument is able to differentiate users on affinity for technology interaction, and its psychometric properties are in accordance with contemporary standards.

The other implication is related to procedures of psychometric validation. The authors conducted most tests in the current study in accordance with Dima's (2018) psychometric protocol, which combines CTT and IRT methods of psychometric analysis and gives the possibility for multi-method assessment. Our decision to use KMO and Barlett's tests for checking EFA assumptions, as well as to prefer R package nFactor for factor retention, were informed by the guidelines for EFA (Field et al., 2012; Howard, 2016). We also tested multivariate normality in accordance with recommendations for CFA (Goodboy & Kline, 2017) and chose R package QuantPsyc (Fletcher, 2012) for this purpose, which was proved to give unbiased results for Mardia's test of multivariate normality (Joenssen & Vogel, 2012). In further research, it might be useful to continue applying multi-method psychometric assessment of instruments and elaborating psychometric protocols in accordance with the most rigorous methodological standards. The extended protocol developed by Dima (2018) was shown to be a useful tool for psychometric assessment, and our amendments to it might be helpful for other researchers.

5.3. Limitations and further research

One of the limitations of the study was related to the sample size determined by the availability of data. Straat, van der Ark, and Sijtsma (2014) showed that even with high item quality, AISP algorithm requires 250 to 500 respondents, while there were only 240 cases in the dataset. It would be recommendable to interpret the results of MSA with caution; in our case, though, they were supported by results of other methods: the item in question (ati03R) had the lowest correlations with other items and the lowest factor loading, and its removal would increase reliability of the scale. It should be stressed, however, that "any reported gains in the reliability of alpha by deleting the item are not representative of the effect this will have on the "true" or population reliability of the scale" (Dunn et al., 2014, p. 403); therefore, based on the results of extended procedures, it can be recommended to further explore the item functioning on a larger sample, and only then final conclusions can be made.

The other limitation is related to inevitable incompleteness of methods applied. As there was no data allowing for analysis of predictive validity of the ATI scale or test-retest reliability, these procedures were left for further research. Violated assumption of multivariate normality, a

Table 4
Item analysis of the ATI scale.

Items	ati01	ati02	ati03R	ati04	ati05	ati06R	ati07	ati08R	ati09
Item-total	.71	.77	.27	.73	.79	.71	.76	.66	.74
Alpha	.89	.88	.92	.89	.88	.89	.88	.89	.89

possible reason of the insufficient CFA fit (Schmitt, 2011), also informed our decision not to include SEM in the set of validation procedures. In the future, thorough check of assumptions and use of SEM for exploring construct validity, in case the assumptions are met, could be beneficial. Some procedures from Dima's (2018) protocol were not conducted in the current study: for instance, the authors decided not to repeat analysis of floor and ceiling effects already presented by Franke et al. (2018). Others, e.g. plotting Item Characteristic Curves, were conducted, as can be seen from the R script, but not included in the paper, as explanations of these methods and presentation of their results would unnecessarily lengthen the paper. The authors are not under the mistaken impression that their psychometric analysis of the scale is in any sense complete but consider it a step in the ongoing validation process.

6. Conclusion

To measure individual aspects of users' interaction with technology, valid and reliable instruments are of paramount importance. As Schimmack (2019) suggested, "the 2020s may become the decade of validation" (p. 5), which should be an iterative process conducted via multi-method assessment.

In this paper, a multi-method approach was applied to the ATI scale (Franke et al., 2018), which was shown to be a valid and reliable instrument recommendable for research in the area of human-technology interaction, and a potential area for further research and improvement of the scale was indicated. Our analysis included hierarchical clustering and methods of CTT and IRT from an extended psychometric protocol recently developed by Dima (2018); suggestions for development of the protocol were made that included more rigorous assumption testing for CFA and EFA and recommendations on the choice of R packages for factor retention and multivariate normality test.

We hope that results of this study will be useful both for researchers who are choosing an instrument to measure users' engagement in technology interaction and for psychometricians constantly widening the repertoire of their methods. Multi-method validation procedures will make research in human-technology interaction more replicable, and its results more implementable for enhancing the effectiveness of human-technology interaction.

Declaration competing of interest

All authors declare that they have no conflicts of interests.

Acknowledgements

We are grateful to Thomas Franke, Daniel Wessel and Christiane Attig, who eagerly shared their data, gave us permission to publish our results and commented on the first draft. We also thank Manuel Prinz for comments that greatly improved the manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.chbr.2020.100004.

References

Attig, C., Wessel, D., & Franke, T. (2017). Assessing personality differences in human-technology interaction: An overview of key self-report scales to predict successful interaction. In C. Stephanidis (Ed.), *HCI international 2017 – posters' extended abstracts* (Vol. 713, pp. 19–29). https://doi.org/10.1007/978-3-319-58750-9_3.

Bakker, M., & Wicherts, J. M. (2014). Outlier removal and the relation with reporting errors and quality of psychological research. *PloS One, 9*(7), 1–9.

Bargas-Avila, J. A., & Hornbæk, K. (2011). Old wine in new bottles or novel challenges. In *Proceedings of the 2011 annual conference on Human factors in computing systems - CHI '11* (pp. 2689–2698). ACM Press.

Boateng, G. O., Neilands, T. B., Frongillo, E. A., Melgar-Quiñonez, H. R., & Young, S. L. (2018). Best practices for developing and validating scales for health, social, and behavioral research: A primer. *Frontiers in Public Health, 6*, 149. https://doi.org/10.3389/fpubh.2018.00149.

Cacioppo, J. T., Petty, R. E., Feinstein, J. A., & Jarvis, W. B. G. (1996). Dispositional differences in cognitive motivation: The life and times of individuals varying in need for cognition. *Psychological Bulletin, 119*(2), 197–253.

Cappelleri, J. C., Jason Lundy, J., & Hays, R. D. (2014). Overview of classical test theory and item response theory for the quantitative assessment of items in developing patient-reported outcomes measures. *Clinical Therapeutics, 36*(5), 648–662. https://doi.org/10.1016/j.clinthera.2014.04.006.

Cazan, A.-M., & Indreica, S. E. (2014). Need for cognition and approaches to learning among university students. *Procedia - Social and Behavioral Sciences, 127*, 134–138. https://doi.org/10.1016/j.sbspro.2014.03.227.

Chavent, M., Kuentz, V., Liquet, B., & Saracco, L. (2011). ClustOfVar: An R package for the clustering of variables. ArXiv:1112.0295 [Stat]. Retrieved from http://arxiv.org/abs/1112.0295.

Christensen, R., & Knezek, G. (2009). Construct validity for the teachers attitudes toward computers questionnaire. *Journal of Computing in Teacher Education, 25*(4), 143–155.

Çokluk, Ö., Büyüköztürk, S., & Kayri, M. (2010). Analyzing factor structure of the scales by hierarchical clustering analysis: An alternative approach. *Australian Journal of Basic and Applied Sciences, 4*, 6397–6403.

Crutzen, R., & Peters, G.-J. Y. (2017). Scale quality: Alpha is an inadequate estimate and factor-analytic evidence is needed first of all. *Health Psychology Review, 11*(3), 242–247. https://doi.org/10.1080/17437199.2015.1124240.

Dima, A. L. (2018). Scale validation in applied health research: Tutorial for a 6-step R-based psychometrics protocol. *Health Psychology and Behavioral Medicine, 6*(1), 136–161. https://doi.org/10.1080/21642850.2018.1472602.

Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology, 105*(3), 399–412. https://doi.org/10.1111/bjop.12046.

Edison, S. W., & Geissler, G. L. (2003). Measuring attitudes towards general technology: Antecedents, hypotheses and scale development. *Journal of Targeting, Measurement and Analysis for Marketing, 12*(2), 137–156. https://doi.org/10.1057/palgrave.jt.5740104.

Emons, W. H. M., Sijtsma, K., & Pedersen, S. S. (2012). Dimensionality of the Hospital Anxiety and Depression Scale (HADS) in cardiac patients. *Assessment, 19*(3), 337–353. https://doi.org/10.1177/1073191110384951.

Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using R*. Los Angeles: SAGE.

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological and Personality Science, 8*(4), 370–378. https://doi.org/10.1177/1948550617693063.

Fletcher, T. D. (2012). QuantPsyc: Quantitative psychology tools. R package, version 1.5. https://CRAN.R-project.org/package=QuantPsyc.

Franke, T., Attig, C., & Wessel, D. (2018). A personal resource for technology interaction: Development and validation of the affinity for technology interaction (ATI) scale. *International Journal of Human-Computer Interaction, 35*(6), 456–467. https://doi.org/10.1080/10447318.2018.1456150.

Franke, T., Attig, C., & Wessel, D. (2020). Data for reanalysis [Data file and codebook]. Retrieved from https://ati-scale.org/wp-content/uploads/2020/01/ATI-MTurk-Data.zip .

Fulford, C. P., & Zhang, S. (1993). Perceptions of Interaction: The critical predictor in distance education. *American Journal of Distance Education, 7*(3), 8–21. https://doi.org/10.1080/08923649309526830.

Goodboy, A. K., & Kline, R. B. (2017). Statistical and practical concerns with published communication research featuring structural equation modeling. *Communication Research Reports, 34*(1), 68–77.

Hogan, T. P., & Agnello, J. (2004). An empirical study of reporting practices concerning measurement validity. *Educational and Psychological Measurement, 64*(5), 802–812. https://doi.org/10.1177/0013164404264120.

Howard, M. C. (2016). A review of exploratory factor Analysis decisions and overview of current practices: What we are doing and how can we improve? *International Journal of Human-Computer Interaction, 32*(1), 51–62. https://doi.org/10.1080/10447318.2015.1087664.

Huang, H.-M., Rauch, U., & Liaw, S.-S. (2010). Investigating learners' attitudes toward virtual reality learning environments: Based on a constructivist approach. *Computers & Education, 55*(3), 1171–1182. https://doi.org/10.1016/j.compedu.2010.05.014.

Hu, L., & Bentler, P. M. (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.

Jackson, D. L., Gillaspy, J. A., & PurcStephenson, R. (2009). Reporting practices in confirmatory factor analysis: An overview and some recommendations. *Psychological Methods, 14*(1), 6–23. https://doi.org/10.1037/a0014694.

Joenssen, D. W., & Vogel, J. (2012). A power study of goodness-of-fit tests for multivariate normality implemented in R. *Journal of Statistical Computation and Simulation*, 1–24. https://doi.org/10.1080/00949655.2012.739620.

Karrer, K., Glaser, C., Clemens, C., & Bruder, C. (2009). Technikaffinität erfassen – der Fragebogen TA-EG [Measuring affinity to technology – the questionnaire TA-EG]. In A. Lichtenstein, C. Stößel, & C. Clemens (Eds.), *Der Mensch im Mittelpunkt technischer Systeme. 8. Berliner Werkstatt Mensch-Maschine-Systeme 7. bis 9. Oktober 2009* (pp. 196–201). Düsseldorf, Germany: VDI.

Korkmaz, S., Goksuluk, D., & Zararsiz, G. (2014). MVN: An r package for assessing multivariate normality. *The R Journal, 6*(2), 151–162.

Lewis, J. R. (2015). Introduction to the special issue on usability and user experience: Psychometrics. *International Journal of Human-Computer Interaction, 31*(8), 481–483. https://doi.org/10.1080/03610918.2015.1064643.

Lezhnina, O. (2020). ExtPsych-ClustATI (Version 1.0.0) [Source code]. http://doi.org/10.5281/zenodo.3660541.

Pruet, P., Ang, C. S., & Farzin, D. (2016). Understanding tablet computer usage among primary school students in underdeveloped areas: Students' technology experience, learning styles and attitudes. *Computers in Human Behavior, 55*, 1131–1144. https://doi.org/10.1016/j.chb.2014.09.063.

R Core Team. (2013). R: A language and environment for statistical computing. Vienna: Austria. Retrieved from http://www.R-project.org.

Raiche, G. (2010). *nFactors: an R package for parallel analysis and non-graphical solutions to the Cattell scree test.R package version 2.3.3*.

Revelle, W. (1978). Iclust: A cluster analytic approach to exploratory and confirmatory scale construction. *Behavior Research Methods & Instrumentation, 10*(5), 739–742. https://doi.org/10.3758/BF03205389.

Revelle, W. (2018). *psych: Procedures for personality and psychological research*. Evanston, Illinois, USA: Northwestern University. https://CRAN.R-project.org/package=psych Version=1.8.12.

Schimmack, U. (2019). The validation crisis in psychology. https://doi.org/10.31234/osf.io/q247m.

Schmettow, M., Noordzij, M. L., & Mundt, M. (2013). An implicit test of UX: Individuals differ in what they associate with computers. In *CHI '13 2013 ACM SIGCHI on human factors in computing systems, Paris, France* (pp. 2039–2048). New York, NY: ACM. https://doi.org/10.1145/2468356.2468722. April 27–May 2.

Schmitt, T. A. (2011). Current methodological considerations in exploratory and confirmatory factor Analysis. *Journal of Psychoeducational Assessment, 29*(4), 304–321. https://doi.org/10.1177/0734282911406653.

Schwab, C. G. G., Dichter, M. N., & Berwig, M. (2018). Item distribution, internal consistency, and structural validity of the German version of the DEMQOL and DEMQOL–proxy. *BMC Geriatrics, 18*(1). https://doi.org/10.1186/s12877-018-0930-0.

Singh, J. (2004). Tackling measurement problems with item response theory. *Journal of Business Research, 57*(2), 184–208. https://doi.org/10.1016/S0148-2963(01)00302-2.

Stochl, J., Jones, P. B., & Croudace, T. J. (2012). Mokken scale analysis of mental health and well-being questionnaire item responses: A non-parametric IRT method in empirical research for applied health researchers. *BMC Medical Research Methodology, 12*(1). https://doi.org/10.1186/1471-2288-12-74.

Straat, J. H., van der Ark, L. A., & Sijtsma, K. (2014). Minimum sample size requirements for mokken scale Analysis. *Educational and Psychological Measurement, 74*(5), 809–822. https://doi.org/10.1177/0013164414529793.

Wang, W., Guedj, M., Bertrand, V., Foucquier, J., Jouve, E., Commenges, D., et al. (2017). A rasch analysis of the charcot-marie-tooth neuropathy score (CMTNS) in a cohort of charcot-marie-tooth Type 1A patients. *PloS One, 12*(1), e0169878. https://doi.org/10.1371/journal.pone.0169878.

Watson, R., van der Ark, L. A., Lin, L.-C., Fieo, R., Deary, I. J., & Meijer, R. R. (2012). Item response theory: How mokken scaling can be used in clinical practice: *Item response theory*. *Journal of Clinical Nursing, 21*(19pt20), 2736–2746. https://doi.org/10.1111/j.1365-2702.2011.03893.x.

Wickham, H. (2017). tidyverse: Easily install and load the 'tidyverse'. R package version 1.2.1. https://CRAN.R-project.org/package=tidyverse.

Young, B. J. (2000). Gender differences in student attitudes toward computers. *Journal of Research on Computing in Education, 33*(2), 204. https://doi.org/10.1080/08886504.2000.10782310.