Thomas Groß*

Validity and Reliability of the Scale Internet Users' Information Privacy Concerns (IUIPC)

Abstract: Internet Users' Information Privacy Concerns (IUIPC-10) is one of the most endorsed privacy concern scales. It is widely used in the evaluation of human factors of PETs and the investigation of the privacy paradox. Even though its predecessor Concern For Information Privacy (CFIP) has been evaluated independently and the instrument itself seen some scrutiny, we are still missing a dedicated confirmation of IUIPC-10, itself. We aim at closing this gap by systematically analyzing IUIPC's construct validity and reliability. We obtained three mutually independent samples with a total of $N = 1031$ participants. We conducted a confirmatory factor analysis (CFA) on our main sample to assert the validity and reliability of IUIPC-10. Having found weaknesses, we proposed a respecified instrument IUIPC-8 with improved psychometric properties. Finally, we confirmed our findings on a validation sample. While we found sound foundations for content validity and could confirm the overall three-dimensionality of IUIPC-10, we observed evidence of biases in the question wording and found that IUIPC-10 consistently missed the mark in evaluations of construct validity and reliability, calling into question the unidimensionality of its sub-scales Awareness and Control. Our respecified scale IUIPC-8 offers a statistically significantly better model and outperforms IUIPC-10's construct validity and reliability. The disconfirming evidence on IUIPC-10's construct validity raises doubts how well it measures the latent variable Information Privacy Concern. The less than desired reliability could yield spurious and erratic results as well as attenuate relations with other latent variables, such as behavior. Thereby, the instrument could confound studies of human factors of PETs or the privacy paradox, in general.

Keywords: IUIPC, privacy concern, measurement instrument, validity, reliability, factor structure

DOI 10.2478/popets-2021-0026
Received 2020-08-31; revised 2020-12-15; accepted 2020-12-16.

***Corresponding Author: Thomas Groß:** Newcastle University, E-mail: thomas.gross@newcastle.ac.uk

1 Introduction

Sound measurement instruments are a key ingredient in the investigation of privacy concern and its impact on human behavior. They act as a measuring stick for privacy concern itself as well as a foundational component for substantive composite models. Thereby, they are a crucial keystone in evaluating human factors of PETs and studying the privacy paradox.

While there has been a diversification of instruments of privacy concern and behaviors [6, 7, 9, 30, 34, 40, 43] also documented in systematic reviews on the privacy paradox [13, 26], Internet Users' Information Privacy Concerns (IUIPC) [30] stands out as a widely adopted scale diligently created in an evolutionary fashion and with a sound theoretical underpinning.

IUIPC is based on Concerns for Information Privacy (CFIP) [40], itself a popular scale measuring organizational information privacy concern, which has been validated in independent empirical studies [18, 41]. Both CFIP and IUIPC scales have been endorsed by Preibusch [34] as sound instruments.

We are interested in IUIPC-10, a 10-item privacy concern scale with the three dimensions Control, Awareness, and Collection. Our interest is rooted in its strong pedigree and its wide-spread use in the investigation of human factors of PETs and of the privacy paradox. While it has seen some scrutiny as part of other studies [32, 38] and questions of its validity have become apparent, there has not yet been a dedicated confirmatory factor analysis to assess its validity and reliability.

We aim at two complementary research questions: (i) First, we investigate to what extent the the validity and reliability of IUIPC-10 can be confirmed. (ii) Second, we consider under which circumstances IUIPC-10 can be employed most reliably, considering the estimation method used. The latter line of inquiry is motivated by design decisions made by Malhotra et al. [30], which are at odds with contemporary recommendations for a sound CFA methodology [5, 11, 24]. Hence, we thereby aim at ruling out possible confounders of our direct replication and at offering empirically grounded recommendations derived from our conceptual replications on how to best use IUIPC.

Our Contributions.
To the best of our knowledge, we established the first dedicated adequately-sized registered independent confirmatory factor analysis of IUIPC-10. We, thereby, offer the first comprehensive disconfirming evidence of the construct validity and reliability of this scale, with wide implications for studies measuring information privacy concern in their endeavor to evaluate the users attitude to PETs or to study the privacy paradox overall. While we found sound foundations in content validity and could confirm the overall three-dimensionality of the scale, we found indications of biases in its questionnaire wording and weaknesses in factorial and convergent validity as well as reliability, especially rooted the sub-scales Control and Awareness. Those weaknesses appeared consistently across our independent samples and irrespective of CFA estimators used. We propose a respecified scale IUIPC-8 that consistently offers a statistically significantly better fit, stronger construct validity and reliability. In terms of analysis methodology, we build a bridge between replicating the exact design decisions of Malhotra et al. [30] to factor analyses especially adept on non-normal, ordinal data following contemporary recommendations [25].

2 Background

2.1 Information Privacy Concern

In defining *information privacy concern* we focus on the conceptual framework of IUIPC. Malhotra et al. [30, p. 337] refer to Westin's definition of information privacy as a foundation of their understanding of privacy concern: "the claim of individuals, groups, or institutions to determine for them selves when, how, and to what extent information about them is communicated to others." Information privacy concern is then defined as "an individual's subjective views of fairness within the context of information privacy."

This framing of information privacy concern is well aligned with the interdisciplinary review of privacy studies by Smith et al. [39], which considered privacy concern as the central antecedent of related behavior in their privacy macro model. Of course, the causal impact of privacy concern on behavior has been under considerable scrutiny with the observation of the privacy attitude-behavior dichotomy—the *privacy paradox* [13]. The intense inquiry of the privacy community into the paradox calls for measuring information privacy concern accurately and reliably. This conviction is rooted in the fact that measurement errors could confound the assessment of users' privacy concern and, thereby, yield an alternative explanation for the privacy paradox: If one does not actually measure privacy concern reliably, it is hardly expected to align with exhibited behavior.

There has been a proliferation of related and distinct instruments for measuring information privacy concern. As a comprehensive comparison would be beyond the scope of this study, we refer to Preibusch's excellent guide to measuring privacy concern [34] for an overview of the field and shall focus on specific comparisons to IUIPC itself. First, we mention *Concern for information privacy* (CFIP) [40] a major influence on IUIPC. It consists of four dimensions—Collection, Unauthorized Secondary Use, Improper Access and Errors. While both questionnaires share questions, CFIP focuses on individuals' concerns about organizational privacy practices and the organization's responsibilities, IUIPC shifts this focus to Internet users framed as consumers and their perception of fairness and justice in the context of information privacy and online companies.

Internet Privacy Concerns (IPC) [9] considered internet privacy concerns with antecedents of perceived vulnerability and control, antecedents familiar from the Protection Motivation Theory (PMT). In terms of the core scale of privacy concern, Dinev and Hart identified two factors (i) Abuse (concern about misuse of information submitted on the Internet) and (ii) Finding (concern about being observed and specific private information being found out). IPC differs from IUIPC in its focus on misuse rather than just collection of information and of concerns of surveillance.

Buchanan et al.'s *Online Privacy Concern and Protection for Use on the Internet* (OPC) [7] measure considered three sub-scales—General Caution, Technical Protection (both on behaviors), and Privacy Attitude. Compared to IUIPC, OPC sports a strong focus on item stems eliciting being concerned and on measures through a range of concrete privacy risks. The authors considered concurrent validity with IUIPC, observing a correlation of $r = .246$ between OPC's privacy concern and the total IUIPC score.

CFIP, IPC and OPC have in common that—unlike IUIPC—they do not explicitly mention the loaded word "privacy."

2.1.1 Genesis of IUIPC

The scale *Internet Users' Information Privacy Concern* (IUIPC) was developed by Malhotra et al. [30], by predominately adapting questions of the earlier 15-item scale Concern for Information Privacy (CFIP) by Smith et al. [40] and by framing the questionnaire for Internet users. CFIP received independent empirical confirmations of its factor structure, first by Stewart and Segars [41], but also by Harborth and Pape [18] on its German translation.

Malhotra et al. [30, pp. 338] conceived IUIPC-10 as a second-order reflective scale of *information privacy concern*, with the dimensions Control, Awareness, and Collection. The authors considered the "act of collection, whether it is legal or illegal," as the starting point of information privacy concerns. The sub-scale Control is founded on the conviction that "individuals view procedures as fair when they are vested with control of the procedures." Finally, they considered being "informed about data collection and other issues" as central concept to the sub-scale Awareness. The authors developed IUIPC in exploratory and confirmatory factor analysis, which we shall review systematically in Section 6.

2.1.2 The Role of Privacy Concern Scales in the Investigation of PETs

The role of privacy concern scales in the investigation of the privacy paradox was well documented in the Systematic Literature Review by Gerber et al. [13]: More than a dozen studies used privacy concern as variable. For instance, Schwaig et al. [37] used IUIPC as instrument in their privacy paradox study.

On human factors of PETs, we have the, e.g., technology acceptance of Tor/JonDonym [21] and anonymous credentials [4]. While these two studies used "perceived anonymity" as a three-item scale, subsequent work by Harborth and Pape used IUIPC as privacy concern scale to evaluate JonDonym[19] and Tor [20].

Furthermore, IUIPC has not only been used in the narrow sense of evaluating the privacy paradox or PETs. Let us highlight a few examples in PoPETS: Pu and Grossklags [35] used a scale adopted from IUIPC Collection (or one of its predecessors) to measure own and friends' privacy concern in their study social app users' valuation of interdependent privacy. Gerber et al. [14] used IUIPC to contextualize their investigation on privacy risk perception. Barbosa et al. [3] used IUIPC as main privacy measure to predict changes in smart home device use.

Given that Smith et al.'s privacy macro model [39] centered on privacy concern as antecedent of behavior and that Preibusch [34] recommended IUIPC as a "safe bet," it is not surprising that IUIPC is high on the list of instruments to analyze privacy paradox or PETs. Thereby, we would expect prolific use of IUIPC in future privacy research and perceive a strong need to substantiate the evidence of its construct validity and reliability.

2.2 Validity and Reliability

When evaluating privacy concern instruments, the dual key questions for privacy researchers interested in the investigation of the human factors of PETs and the privacy paradox are: (i) Are we measuring the hidden latent construct Privacy concern accurately? (validity) (ii) Are we measuring privacy concern consistently and with an adequate signal-to-noise ratio? (reliability)

Validity refers to whether an instrument measures what it purports to measure. Messick offered an early well-regarded definition of validity as the "integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores" [31]. Validity is inferred—judged in degrees—not measured. In our work, we take a pragmatic empiricist's approach focusing on the validation procedure and evidence.

2.2.1 Content Validity

Content validity refers to the relevance and representativeness of the content of the instrument, typically assessed by expert judgment. We shall evaluate content validity together in keeping with evidence on the craft of the questionnaire design, incl. question format, language used, question order, and, hence, assess *psychometric barriers* in the form of biases rooted in the questionnaire wording [33, p. 128].

Priming: *Priming* means that mentioning a concept activates it in respondents minds and makes it more easily accessible in subsequent questions. Priming is, for instance, created by the use of a *loaded word* such as "security." It can invoke the respondents' *social desirability bias*.

Leading Questions: *Leading questions* elicit agreement or specific instances of a general term, introduce a bias towards that lead.

Double-Barreled Questions: *Double-barreled questions* consist of two or more questions or clauses, making it difficult for the respondent to decide what to answer to, causing nondifferentiated responses.

Positively-Oriented Questions: Exclusively using positively framed wording for questions leads to *nondifferentiation*/straightlining and, thereby, to accommodating the *acquiescent response bias*.

2.2.2 Construct Validity

First, we seek evidence of *factorial validity*, that is, evidence that that factor composition and dimensionality are sound. While IUIPC is a *multidimensional* scale with three correlated designated dimensions, we require *unidimensionality* of each sub-scale, a requirement discussed at length by Gerbing and Anderson [15]. The empirical evidence for factorial validity is the found in the adequacy of the hypothesized model's fit and passing the corresponding fit hypotheses of a confirmatory factor analysis for the designated factor structure [2, 15, 25]. Specifically, we prioritize the following fit metrics and hypotheses, refering the interested reader to Appendix C for further explanations:

Goodness-of-Fit χ^2: Measures the exact fit of a model and gives rise to the accept-support exact-fit test against null hypothesis $H_{\chi^2,0}$.

RMSEA: Root Mean Square Estimate of Approximation, an absolute badness-of-fit measure estimated as $\hat{\varepsilon}$ with its 90% confidence interval, yielding a range of fit-tests: close fit, not-close fit, and poor fit with decreasing tightness requirements.

Further common criteria, such as CFI or SRMR, are defined in Appendix C.

Convergent validity [17, pp. 675] (convergent coherence) on an item-construct level means that items belonging together, that is, to the same construct, should be observed as related to each other. Similarly, *discriminant validity* [17, pp. 676] (discriminant distinctiveness) means that items not belonging together, that is, not belonging to the same construct, should be observed as not related to each other. Similarly, on a sub-scale level, we expect factors of the same higher-order construct to relate to each other and, on hierarchical factor level, we expect all 1st-order factors to load strongly on the 2nd-order factor.

While a poor local fit and tell-tale residual patterns yield disconfirming evidence for convergent and discriminant validity, further evidence is found evidence inter-item correlation matrices. We expect items belonging to the same sub-scale to be highly correlated (converge on the same construct). At the same time correlation to items of other sub-scales should be low, especially lower than the in-construct correlations [25, pp. 196].

These judgments are substantiated with empirical criteria, where we highlight the following metrics:

Standardized Factor Loadings β: Z-transformed factor scores, typically reported in factor analysis.

Variance Extracted R^2: The factor variance accounted for, giving rise to the Average Variance Extracted (AVE) defined subsequently in reliability Section 2.2.3.

Heterotrait-Monotrait Ratio (HTMT): The *Heterotrait-Monotrait Ratio* is the ratio of the the avg. correlations of indicators across constructs measuring different phenomena to the avg. correlations of indicators within the same construct [22].

We gain empirical evidence in favor of convergent validity [17, pp. 675] (i) if the variance extracted by an item $R^2 > .50$ entailing that the standardized factor loading are significant and $\beta > .70$, and (ii) if the internal consistency (defined in Section 2.2.3) is sufficient ($AVE > .50$, $\omega > AVE$, and $\omega > .70$). The analysis yields empirical evidence of discriminant validity [17, pp. 676] (i) if the square root of AVE of a latent variable is greater than the max correlation with any other latent variable (Fornell-Larcker criterion [12]), (ii) if the Heterotrait-Monotrait Ratio (HTMT) is less than .85 [1, 22].

2.2.3 Reliability

Reliability is the extent to which a variable is consistent in what is being measured [17, p. 123]. It can further be understood as the capacity of "separating signal from noise" [36, p. 709], quantified by the ratio of true score to observed score variance. [25, pp. 90] We evaluate *internal consistency* as a means to estimate reliability from a single test application. Internal consistency entails that items that purport to measure the same construct produce similar scores [25, p. 91]. We will use the following internal consistency measures:

Cronbach's α: Is based on the average inter-item correlations.

Congeneric Reliability ω: The amount of general factor saturation (also called *composite reliabil-*

ity [25, pp. 313] or construct reliability (CR) [17, p. 676] depending on the source).

AVE: *Average Variance Extracted* (AVE) [25, pp. 313] is the average of the squared standardized loadings of indicators belonging to the same factor.

Thresholds for reliability estimates like Cronbach's α or Composite Reliability ω are debated in the field, where many recommendations are based on Nunnally's original treatment of the subject, but equally often misstated [27]. The often quoted $\alpha \geq .70$ was described by Nunnally only to "save time and energy," whereas a greater threshold of .80 was endorsed for basic research [27]. While that would be beneficial for privacy research as well, we shall adopt reliability metrics $\alpha, \omega \geq .70$ as suggested by Hair et al. [17, p. 676]. We further require $AVE > .50$.

2.3 Factor Analysis as Validation Tool

Factor analysis is an excellent tool to establish construct validity and reliability of an instrument. We introduce factor analysis concepts in the Appendix C and assume knowledge of the standard factor analysis with Maximum Likelihood estimation. Here we shall only mention properties of estimators MLM and robust WLS, developed to handle non-normal, ordinal data.

A maximum-likelihood estimation with robust standard errors and Satorra-Bentler scaled test statistic (MLM) is robust against some deviations from normality [24, p. 122]. Lei and Wu [28, p. 172] submit that Likert items with more than five points approximate continuous measurement enough to use MLM; this stance, however, is not universally endorsed [5, 11].

Kline recommends to use of estimators specializing on ordinal data [24, p. 122], such as robust weighted least squares ($WLSMVS$[1]). The following explanation summarizes Kline [25, pp. 324]. In general, robust WLS methods associate each indicator with a latent response variable for which the method estimates thresholds that relate the ordinal responses on the indicator to a continuous distribution on the indicator's latent response variable. Then, the measurement model is evaluated with the latent response variables as indicators, while optimizing to approximate the observed polychoric correlations between the original indicator variables. In addition, the method will compute robust standard errors and corrected test statistics. Because the method also needs to estimate the thresholds, the number of free parameters will be greater than in a comparable ML estimation. Reported estimates and R^2 relate to the latent response variable, not the original indicators themselves.

Because of the level of indirection and the estimation of non-linearly related thresholds, robust WLS is a quite different kettle of fish than Maximum Likelihood estimation on assumedly continuous variables: they are not easily compared by trivially examining fit indices. The estimates of robust WLS need to be interpreted differently than in ML: they are the *probit* of individuals' response to an indicator instead of a linear change in an indicator [5]. The thresholds show what factor score is necessary for the respective option of an indicator or higher to be selected with 50% probability.

The advantages and disadvantages of robust WLS have been carefully evaluated in simulation studies with known ground truth [10, 29]: (i) Robust WLS models show little bias in the parameter estimation even es the level of skewness and kurtosis increased. Unlike ML models, do not suffer from out-of-bounds estimations of indicator variables. (ii) There is contradictory evidence on standard errors, where robust WLS may be subject to greater amounts of bias. (iii) χ^2 fit indices of robust WLS are inflated for larger models or smaller samples. (iv) Robust WLS may inflate correlation estimates.

3 Related Work

Sipior et al. [38] observed that IUIPC was under-studied at the time, offered a considerate review of the related literature, and focused their lens on the role of trusting and risk beliefs in IUIPC's causal model. With the caveat of being executed on a small sample of $N = 63$ students, their research could "not confirm the use of the IUIPC construct to measure information privacy concerns." Even though they also excluded one Control item, their concerns, however, were mostly focused on the causal structure of IUIPC and not on the soundness of the underlying measurement model of IUIPC-10. Our study goes beyond Sipior et al.'s by evaluating the heart of IUIPC, that is, its measurement model, and by doing so in adequately sized confirmatory factor analyses.

Morton [32, p. 472] conducted an adequately sized exploratory factor analysis of IUIPC-10 ($N = 353$) as part of his pilot study for the development of the scale on Dispositional Privacy Concern (DPC). He observed a misloading of the item we call awa3 between the di-

1 weighted least squares with robust standard errors and a Satterthwaite mean- and variance-adjusted test statistic

mensions Awareness and Control in that EFA. He chose to exclude from IUIPC-10 the two items we call ctrl3 and awa3. Even though not highlighted as a main point of the paper, Morton's EFA raised concerns on the validity and factor structure of IUIPC-10. Our analysis differs from Morton's by employing confirmatory factor analysis poised to systematically establish construct validity, by offering a wide range of diagnostics beyond the factor loadings and by re-confirming our analyses on an independent validation sample. While Morton's pilot EFA largely yields statements on his sample, our analysis is a pre-registered confirmatory study yields at results holding generally for the instrument itself and irrespective of estimation methods used.

To the best of our knowledge, the validity and reliability of IUIPC-10 have not undergone an adequately sized dedicated independent analysis, to date. As a starting point for that inquiry, we offer a detailed evaluation of the original IUIPC-10 scale in Section 6.

4 Aims

4.1 Conception and Content Validity

We are interested in the roots of IUIPC, especially its content validity and its reported psychometric properties at its conception.

RQ 1 (Content Validity of IUIPC)**.** *What are the qualitative properties of IUIPC's content validity, that is, relevance and representativeness of the content in the instrument.*

4.2 Construct Validity and Reliability

Our main goal is an independent confirmation of the IUIPC-10 instrument by Malhotra et al. [30], where we focus on construct validity and reliability.

RQ 2 (Confirmation of IUIPC)**.** *To what extent can we confirm IUIPC-10's construct validity and reliability?*

This aim largely entails confirming the factorial validity, that is, the three-dimensional factor structure, of IUIPC. The first inquiry there is to compare alternative models of the IUIPC with different factor solutions.

Second, we will gather further evidence for factorial validity by seeking to fit IUIPC-10 hypothesized second-order model, where the unidimensionality of its sub-scales, that is, the absence of cross-loadings is a key consideration. This will be tested with statistical inferences on the models global fit based on the statistical hypotheses introduced in Appendix C indicating increasingly worse approximations: (i) Exact Fit ($H_{\chi^2,0}$), (ii) Close Fit ($H_{\varepsilon_0 \leq .05,0}$), (iii) Not Close Fit ($H_{\varepsilon_0 \geq .05,0}$), (iv) Poor Fit ($H_{\varepsilon_0 \geq .10,0}$). We further evaluated the combination rule used by Malhotra et al. [30]: (i) CFI $> .95$, (ii) GFI $> .90$, (iii) RMSEA $< .06$. This global fit analysis will be complemented by an assessment of local fit on residuals.

This inquiry is complemented by analyses of convergent and discriminant validity on the criteria established in Section 2.2.2 similarly to analyses of internal-consistency reliability on criteria from Section 2.2.3

Overall, the aim of evaluating the construct validity and reliability of IUIPC-10 is not just about a binary judgment, but a fine-grained diagnosis of possible problems and viable improvements. This wealth of evidence to enable privacy researchers to form their own opinions.

4.3 Estimator Appraisal

As second line of inquiry, we considered multiple estimation methods in conceptual replications shown in Figure 1 on the horizontal axis.

RQ 3 (Estimator Invariance)**.** *To what extent do the confirmation results in regards to RQ 2 hold irrespective of the estimator used?*

For that, we expect the statistical hypotheses of RQ 2 to yield the same outcome across estimation methods. For respecifications, we expect the fit indices (especially CFI and CAIC) to show appreciable improvements comparing the models on their respective estimators shown in Figure 1 on the vertical axis.

In addition, we aim at gaining an empirical underpinning to design decisions made in the field with respect to the methodological setup of CFAs and SEMs with IUIPC and similar scales.

RQ 4. *Which estimator is most viable to create models with IUIPC, measured in ordinal 7-point Likert items?*

We aim at investigating the viability of alternatives to the maximum likelihood (ML) estimation: (i) scaled estimation (MLM) and (ii) estimation specializing on ordinal variables (robust WLS). As discussed in Section C, robust WLS is a far cry from ML/MLM estimation.

Hence, a plain comparison on their fit indices, such as on the consistent Akaike Information Criteria (CAIC), may lead us astray: their fit measures are not directly comparable in a fair manner.

Thereby, the question becomes: Are the respective estimations viable in their own right, everything else being equal? To what extent do the estimators offer us a plausible approximation of IUIPC? For these questions we aim at estimating mean structures throughout such that we can assess their first-order predictions on indicators. Without knowing the ground truth of the true IUIPC scores of our samples, the assessment on what estimator is most viable will be largely qualitative.

5 Method

The project was registered on the Open Science Framework[2]. The OSF project contains the registration of the study as well as technical supplementary materials, incl. R covariance matrices and related data needed to reproduce the ML- and MLM-estimated CFAs. There is a comprehensive arXiv report with supplementary materials [16]. The statistics were computed in R largely using the package lavaan, where graphs and tables were largely produced with knitr. The significance level was set to $\alpha = .05$.

5.1 Ethics

The ethical requirements of the host institution were followed and ethics cases registered. Participants were recruited under informed consent. They could withdraw from the study at any point. They were enabled to ask questions about the study to the principal investigator. They agreed to offer their demographics (age, gender, mother tongue) as well as the results of the questionnaires for the study. Participants were paid standard rates for Prolific Academic, £12/hour, which is greater than the UK minimum wage of £8.21 during the study's timeframe. The data of participants was stored on encrypted hard disks, their Prolific ID only used to ensure independence of observations and to arrange payment.

5.2 Sample

We used three independent samples, A, B, V in different stages of the analysis. Auxiliary sample A was collected in a prior study and aimed at a sample size of 200 cases.

Base sample B and validation sample V were collected for a current investigation. They had a designated sample size of 420 each, based on an *a priori* power analysis for structural equation modeling with RMSEA-based significance tests.

While all three samples were recruited on Prolific Academic, B and V were recruited to be representative of the UK census by age and gender. The sampling frame was Prolific users who were registered to be residents of the UK, consisting of 48,454 users at sampling time. The sampling process was as follows: 1. Prolific presented our studies to all users with matching demographics, 2. the users could choose themselves whether they would participate or not. We prepared to enforce sample independence by uniqueness of the participants' Prolific ID.

We planned for excluding observations from the sample, without replacement, because (i) observations were incomplete, (ii) observations were duplicates by Prolific ID, (iii) participants failed more than one attention check, (iv) observations constituted multi-variate outliers determined with a Mahalanobis distance of 12 or greater.

5.3 Analysis Approach

We analyzed Malhotra et al.'s original IUIPC publication [30] wrt. RQ 1 with a qualitative review on content validity and with quantitative assessments of construct validity and reliability evidenced at the time. In absence of the original dataset, this analysis was based on the reported descriptives in the publication.

We bridged between a direct replication of the IUIPC-10 analysis approach and conceptual replications adopting to non-normal, ordinal data. We illustrate the dimensions of our approach in Figure 1. As a direct replication with ML estimation and without distribution or outlier consideration would have exposed this study to unpredictable confounders, we computed our analysis with three estimators *ceteris paribus*. We computed all models including their mean structures.

We faced the didactic challenge that even though robust WLS would be best suited for the task at hand [5], it is least used in the privacy community. As discussed in background Section 2.3, its probit estimation is in-

2 OSF: https://osf.io/5pywm

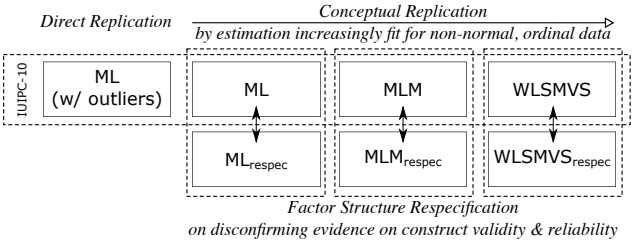


Fig. 1. Analysis approach

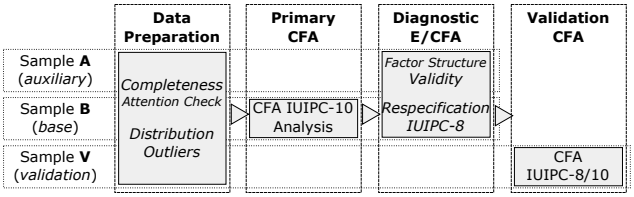


Fig. 2. Which steps were taken on what sample

terpreted differently to more common ML CFA. Hence, we chose to make the MLM model the primary touch stone for our analysis: It carries the advantages of being robust to moderately skewed non-normal data and of yielding interpretations natural to the community.

In our analysis process depicted in Figure 2, we used our three samples deliberately: For the factor analysis of IUIPC-10, we used base Sample B as main dataset to work with. We retained Sample A as an auxiliary sample to conduct exploratory factor analyses and to have respecification proposals informed by more than one dataset, thereby warding against the impact of chance. Sample V was reserved for validation after a final model was chosen. Figure 2 illustrate this relationship of the different samples to analysis stages.

First, we established a sound data preparation, including consideration for measurement level and distribution as well as outliers. Second, we computed a covariance-based confirmatory factor analysis on the IUIPC-10 second-order model [30], complemented with alternative one-factor and two-factor models. This comparison served to confirm the three-dimensionality of IUIPC. We evaluated the hypothesized IUIPC-10 model on Sample B, gathering evidence for construct validity in the form of factorial validity evident in global and local fit, convergent and discriminant validity, as well as reliability.

Having found inconsistencies, we then engaged in a diagnosis and respecification stage. Therein, we also computed a parallel polychoric factor analysis and EFAs on samples A and B to re-assess the three-dimensional factor structure itself and to hunt down patterns of weaknesses. From this evaluation, we prepared a respecified IUIPC-8 which was first evaluated on Sample B. We compared the non-nested models of IUIPC-10 and IUIPC-8 with the Vuong Likelihood Ratio Test [42] on the ML estimation. Otherwise, we compared between fit indices, focusing on an evenly weighted CAIC for non-nested comparisons.

Finally, once respecification and design decisions were settled, we entered the CFA validation stage. Therein, we compared the performance of the original IUIPC-10 and the respecified IUIPC-8 on the independent validation Sample V.

6 Review of IUIPC-10

Internet users' information privacy concerns (IUIPC-10) [30] was created in two studies, determining a preliminary factor structure in an EFA on Study 1 ($N_{\mathsf{IUIPC},1} = 293$) and confirming it in a LISREL covariance-based ML CFA on Study 2 ($N_{\mathsf{IUIPC},2} = 449$). We have asked Malhotra et al. for a dataset or covariance matrix to directly compare against their results. Sung S. Kim [23] was so kind to respond promptly and stated that they could locate these data.

6.1 Content Validity

The authors [30, pp. 338] make a compelling and well-argued case for the content relevance of the information privacy concern dimensions of Collection, Control, and Awareness (cf. in Section 2.1.1). Being rooted in Social Contract (SC) theory, the authors focus on one of the key SC principles they quoted as "norm-generating microsocial contracts must be founded in informed consent, buttressed by rights of exit and voice,"which, in turn, underpins the respondents perception of fairness of information collection contingent on their granted control and awareness of intended use.

The questionnaire consisted of ten Likert 7-point items anchored on 1="Strongly disagree" to 7="Strongly Agree." We included the questionnaire in the Materials Appendix A Table 9. In terms of question format, we observe two types of questions present (i) statements of belief or conviction, e.g., "Consumer control of personal information lies at the heart of consumer privacy." (ctrl2) and (ii) statements of concern, e.g., "It usually bothers me when online companies ask me for personal information." (coll1) We would classify

ctrl1, ctrl2, awa1 and awa2 as belief statements, the remainder as concern statements.

Considering the temporal reference point of the questions, we observe that the questionnaire aims at long-term *trait* statements, evoked for instance by keywords like "usually." Consequently, we believe IUIPC not to respond strongly to respondents short-term changes in *state*.

When it comes to content relating to psychometric barriers and biases, we find that the questionnaire mentions the loaded word "privacy" four times. It further uses loaded words, such as "autonomy." Hence, we expect the questionnaire to exhibit a systematic priming bias and a social desirability bias, leading to a negative skew of measured scores. The priming is aggravated by the question order, in which the loaded words are most predominant in the first three items.

We find leading questions, too: ctrl3 "I believe that online privacy is invaded when control is lost or unwillingly reduced as a result of a marketing transaction," is a leading question towards thinking about the more specific theme of "marketing transactions." Other questions, such as awa1 "Companies seeking information online should disclose the way the data are collected, processed, and used." induce agreement—why would respondent disagree with such a statement?

Two items exhibit a double-barreled structure: ctrl3—"I believe that online privacy is invaded when control (i) is lost or (ii) unwillingly reduced. . . " We find for awa3—"It is very important to me that I am (i) aware and (ii) knowledgeable. . . In both cases, we can ask how a participant will answer if only one of the two clauses is fulfilled, or both.

Finally, we find that the questionnaire only contains positively-oriented items. The absence of reverse-coding may lead to nondifferentiation, a risk also observed by Preibusch [34]. This can set up the respondents' acquiescent response bias.

6.2 Sample

The samples were obtained by "students in a marketing research class at a large southeastern university in the United States [. . .] collecting the survey data" [30, p. 343] from households in the catchment area of the university in one-to-one interviews. No explicit survey population or sampling frame was reported, placing the sampling process in the realm of judgment sampling. While there was no information given how the sample size was determined, the total sample for Study 2 ($N_{\mathsf{IUIPC},2} = 449$) was not unreasonable.

Table 1. Fit of IUIPC-10's second-order model [30, p. 346]. The tests of the exact-fit hypothesis on the χ^2 and the not-close fit hypothesis on the RMSEA failed. The tests of close-fit and poor-fit hypotheses passed. The combination of CFI, GFI, and RMSEA supports a satisfactory fit.

χ^2	*df*	*p*	CFI	GFI	RMSEA	$p_{\epsilon_0 \leq .05}$
73.19	32	< .001	.98	.97	.054 [.037, .070]	.338

Note: $N_{\mathsf{IUIPC}} = 449$. RMSEA with inferred 90% CI.

6.3 Construct Validity

6.3.1 Assumptions

In terms of assumptions and requirements of Maximum Likelihood estimation, the paper did not mention how distribution, univariate and multivariate outliers were handled. Kim [23] clarified that they "did not check the distribution or outliers" and that they "relied on the robustness of maximum likelihood estimation", a case Bovaird and Koziol [5, p. 497] called "ignoring ordinality." In the field, there are practitioners considering the ML estimator robust enough to handle ordinal data with more than five levels as well as empirical analyses cautioning against this practice [5, 11]. We found IUIPC-10 surveys to yield non-normal data with a negative skew throughout, rendering the questionnaire scores less suitable to be covered by ML-robustness results.

6.3.2 Factorial Validity

In terms of global fit as evidence for factorial validity, we outlined the fit measures reported for the original IUIPC instrument [30] in Table 1. Though not reported, we estimated the χ^2 p-value, the RMSEA 90% confidence interval, and $p_{\epsilon_0 \leq .05}$ from the χ^2 test statistic.

In terms of statistical inferences, we observed that the original IUIPC model failed the exact-fit test and the not-close fit test ($\hat{\epsilon}_{\mathsf{UL}} < .05$). It passed the the close-fit test, poor-fit test and the combination rule (HBR). The authors did not report the model's SRMR.

The original IUIPC paper did not contain an analysis of residuals. We did not have the data at our disposal to compute it ourselves [23] and could, thereby, not evaluate the local fit. Hence, the evidence for factorial validity was incomplete.

Table 2. Validity and reliability evidence on IUIPC [30, Tab. 2]. Low AVE with awa shy of the $AVE > .50$ criterion cautions against low internal consistency, even if construct reliability $\omega > .70$ is sufficient, entailing a moderate signal-to-noise ratio S/N_ω. That the $\sqrt{AVE}$ of awa and ctrl on the diagonal of the correlation table is less than the correlation with the respective other factor violates the Fornell-Larcker criterion for discriminant validity.

						Correlations			
		M	SD	AVE	ω	S/N_ω	1	5	6
1.	coll	5.63	1.09	.55	.83	4.88	.74		
5.	awa	6.21	0.87	.50	.74	2.85	.66	.71	
6.	ctrl	5.67	1.06	.54	.78	3.55	.53	.75	.73

Note: Value on the diagonal is the square root of AVE.

6.3.3 Convergent and Discriminant Validity

In the item-level evaluation of convergent validity, we first examined the factor loadings reported for IUIPC. The EFA of IUIPC (of unspecified rotation/transformation) is stated by the authors to have retained items that loaded greater than .70 on their designated factors, and less than .40 on other factors. The CFA of the measurement model was reported to have had a minimal factor loading of .61 for Awareness. Neither a detailed factor loading table was reported, nor standardized loadings or R^2 for individual items. Criteria for AVE and composite reliability were fulfilled, just so for AVE. In terms of discriminant validity, we find that the Fornell-Larcker criterion is violated for Awareness (awa) and Control (ctrl), where the their correlation is greater than the square root of their respective AVE.

6.4 Reliability

Considering the internal consistency criteria vis-à-vis of Table 2, we find that the Average Variance Extracted criterion $AVE > .50$ is just so fulfilled, with awareness on the boundary of acceptable. The composite reliability ω is greater than .70 throughout, with coll achieving the best value (.83). The reported reliability is low for AVE but decent for composite reliability.

6.5 Summary

Concerning RQ 1, content validity is a strong point of IUIPC as the argument on relevance seems compelling. We observed problems in the questionnaire wording that could introduce systematic biases into the scale, though. Considering the two-step process with a preliminary EFA and a subsequent CFA, there is an impression that IUIPC has been diligently done. In terms of construct validity, we found that IUIPC reported a satisfactory global fit, while unchecked assumptions, an estimation at odds with non-normal, ordinal data, and the missing information on local fit weakened the case. While evidence of convergent validity was scarce without a factor loading table or standardized loadings to work with, discriminant validity was counter-indicated. Finally, the low—if not disqualifying—AVE in the reliability inquiry will caution privacy researchers to expect an only moderate signal-to-noise ratio (S/R_ω between 2.85 and 4.88) and attenuation of effects on other variables.

7 Quantitative Results

7.1 Sample

We refined the three samples A, B and V in stages, where Table 3 accounts for the refinement process. First, we removed incomplete cases without replacement. Second, we removed duplicates across samples by the participants' Prolific ID. Third, we removed cases in which participants failed more than one attention check (FailedAC > 1). Overall, of the $N_C = 1074$ complete cases, only 5.3% were removed due to duplicates or failed attention checks.

The demographics the samples are outlined in Table 4. In samples B and V meant to be UK representative, we found a slight under-representation of elderly participants compared to the UK census age distribution.

Table 3. Sample Refinement

Phase	A		B		V	
	Excl.	Size	Excl.	Size	Excl.	Size
Starting Sample		226		473		467
Incomplete	0	226	58	415	34	433
Duplicate	7	219	25	390	0	433
FailedAC> 1	14	205	11	379	0	433
MV Outlier	4	201	9	370	14	419
Final Sample	$N'_A = 201$		$N'_B = 370$		$N'_V = 419$	

Note: $N_A = 205$, $N_B = 379$, $N_V = 433$ are after attention checks.

Table 4. Demographics

(a) Sample A

	Overall
N_A	205
Gender (%)	
Female	80 (39.0)
Male	125 (61.0)
Rather not say	0 (0.0)
Age (%)	
18-24	109 (53.2)
25-34	71 (34.6)
35-44	18 (8.8)
45-54	4 (2.0)
55-64	3 (1.5)
65+	0 (0.0)

(b) Sample B

	Overall
N_B	379
Gender (%)	
Female	197 (52.0)
Male	179 (47.2)
Rather not say	3 (0.8)
Age (%)	
18-24	41 (10.9)
25-34	72 (19.0)
35-44	84 (22.2)
45-54	57 (15.0)
55-64	97 (25.6)
65+	28 (7.4)

(c) Sample V

	Overall
N_V	433
Gender (%)	
Female	217 (50.1)
Male	212 (49.0)
Rather not say	4 (0.9)
Age (%)	
18-24	92 (21.2)
25-34	143 (33.0)
35-44	83 (19.2)
45-54	58 (13.4)
55-64	44 (10.2)
65+	13 (3.0)

Note: Samples B and V were drawn to be representative of the UK census by age and gender; Sample A was not.

7.2 Descriptive Statistics

Following Kline's guidance on distribution assumptions [25, pp. 74], we found all indicator variables of all samples to be substantively negatively skewed, meaning that there are relatively few small values and that the distribution tails off to the left, with the most extreme skew being −2.09. In general, all indicators apart from coll1 showed a substantive positive kurtosis, that is, peakedness, less than 2.4. While this pattern of substantive non-normality was present in the indicator distributions, we also found it in the IUIPC sub-scales and illustrate these distributions in Table 5 and Figure 3. We observed that the three samples had approximately equal distributions by sub-scales.

Our IUIPC-10 samples yielded 5% univariate outliers by the robust outlier labeling rule and 3% multivariate outliers with a Mahalanobis distance of 12 or greater [25, pp. 72].

Regarding the requirements for ML estimation, we find a situation of non-continuous measurement in which multi-variate normality is violated. Our data preparation handled the outliers as recommended.

Table 5. Means (SDs) of the summarized sub-scales of IUIPC-10

	Sample A	Sample B	Sample V	Malhotra et al.
ctrl	5.82 (0.99)	5.93 (0.78)	5.86 (0.84)	5.67 (1.06)
awa	6.22 (0.78)	6.51 (0.52)	6.43 (0.66)	6.21 (0.87)
coll	5.48 (1.12)	5.58 (1.12)	5.60 (1.04)	5.63 (1.09)
iuipc	5.84 (0.75)	6.00 (0.61)	5.96 (0.64)	5.84 (1.01)

7.3 Construct Validity

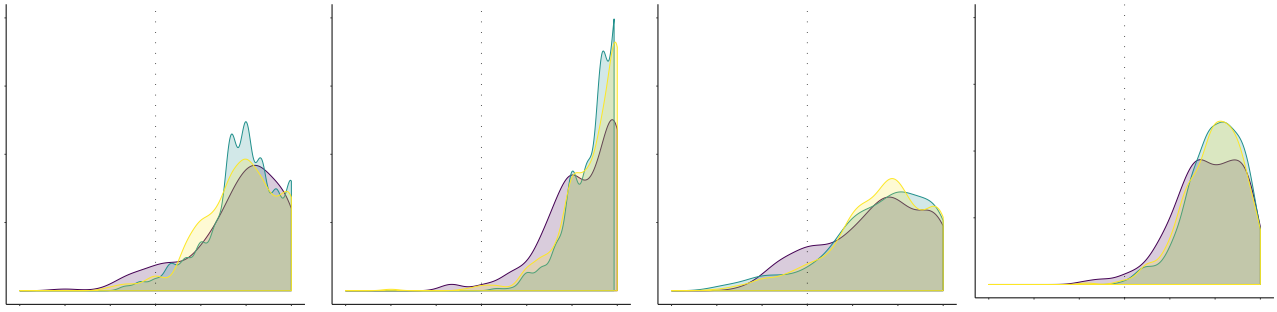
7.3.1 Factorial Validity

To confirm the hypothesized factor structure of IUIPC-10, we computed confirmatory factor analyses on one-factor, two-factor and the hypothesized three-dimensional second-order model. We present the fit of the respective estimations in Table 12. By a likelihood-ratio χ^2 difference test, we concluded that the two-factor solution was statistically significantly better than the one-factor solution, $\chi^2(1) = 138.761, p < .001$. In turn, the three-factor solutions were statistically significantly better than the two-factor solution, $\chi^2(2) = 49.957, p < .001$. We accepted the hypothesized three-factor second-order model, offering confirming evidence for its factorial validity. To further test the construct validity of the three-factor second-order model, we conducted a confirmatory factor analysis of the IUIPC-10 measurement model on Sample B. We included the model's path plot in Figure 5 in Appendix B.

Global Fit

Our first point of call for further evaluating the factorial validity of IUIPC-10 is global fit. We included an overview of the fit measures on different samples in Table 7, drawing attention to the top row.

First, we observed that the exact-fit test failed for IUIPC-10 irrespective of estimator, that is, the exact-fit null hypotheses $H_{\chi^2,0}$ were rejected with the χ^2-tests being statistically signficant. For the RMSEA-



(a) Control **(b)** Awareness **(c)** Collection **(d)** IUIPC-10 Overall

Fig. 3. Density of IUIPC-10 subscale responses across samples (A: violet, B: green, V: yellow). Note: All graphs are on the same scale. The dotted mid-line represents the neutral option of the 7-point Likert scale

based hypotheses we have: (i) The close-fit test evaluating whether RMSEA is likely less or equal .05 failed irrespective of estimator. (ii) The not-close-fit hypothesis could not be rejected for either estimator, withholding support for the models. (iii) Finally, the poor-fit hypothesis could not be rejected either for any models, with the upper bound of the RMSEA CI being greater than or equal as .10, indicating a poor fit.

The fit indices CFI and SRMR yielded .92 and .10 for the ML based models, respectively, not supporting the models. None of the models passed the *HBR* combination rule used by Malhotra et al. The direct replication of IUIPC with ML estimation and outliers present fared more poorly than the corresponding ML models implementing the stated assumptions: $\chi^2(32) = 202.789, p < .001$; CFI=.90; RMSEA=.12 [.10, .13]; SRMR=.11; CAIC=333.8.

Overall, we conclude that the global fit of the model was poor and that we found disconfirming evidence for IUIPC-10's factorial validity. This disconfirmation of the CFA held irrespective of the data preparation and estimator employed. Our further examination of construct validity will be on the touch-stone MLM model.

Local Fit

The correlation residuals showed appreciable patterns of positive correlations greater than .10 with ctrl3 and awa3, matched with statistically significant standardized covariances. This indicated considerable misloading on these indicators and disconfirmed the unidimensionality of the corresponding sub-scales.

7.3.2 Convergent and Discriminant Validity

We first analyzed the standardized loadings and the variance explained in Table 6. Therein, we found that ctrl3 and awa3 only exhibited .17 and .20 of the common factor variance, respectively. Those values were considerably below par ($R^2 > .50$). The corresponding loadings would be classified according to Hair et al. [17, p. 153] as just past minimally acceptable, but not practically significant. In the evaluation of the second-order model, we found the standardized loading of Collection on IUIPC on the low side. In terms of convergent validity, we evaluated the Average Variance Extracted (AVE) and Composite Reliability (CR) ω in Table 6. While the CR ω being greater than the AVE for all three dimensions indicated support, we observed that the AVE was less than .50 for both Control and Awareness and thereby showed that there is little common factor variance extracted on average. Similarly, their $\omega < .70$ implied sub-par convergent validity.

For discriminant validity, we found the Fornell-Larcker and HTMT criteria fulfilled, offering support for the specified models.

7.4 Reliability: Internal Consistency

Let us consider the reliability criteria derived from the MLM CFA model in Table 6. Considering Cronbach's α, we observed estimates for Control and Awareness less than the .70, what Nunnally classified only acceptable to "save time and energy." The Composite Reliability estimate $\omega = .66$ was equally below this just acceptable threshold.

Table 6. Factor loadings and their standardized solution of the MLM CFA of IUIPC-10 on Sample B.
We find sub-par standardized loadings $\beta < .70$ for ctrl3 and awa3, yielding a poor variance extracted $R^2 \leq .20$ and, thereby, low $AVE < .50$ for control and awareness, indicating sub-par internal consistency. The equally sub-par construct reliability $\omega < .70$ yields a low signal-to-noise ratio less than 2.

Factor	Indicator	Factor Loading				Standardized Solution				Reliability				
		λ	SE_λ	Z_λ	p_λ	β	SE_β	Z_β	p_β	R^2	AVE	α	ω	S/N_ω
ctrl	ctrl1	1.00^+				0.73	0.05	15.14	$< .001$	0.54	0.40	0.62	0.66	1.92
	ctrl2	1.00	0.11	8.76	$< .001$	0.73	0.05	13.73	$< .001$	0.53				
	ctrl3	0.59	0.11	5.36	$< .001$	0.41	0.06	6.89	$< .001$	0.17				
aware	awa1	1.00^+				0.74	0.05	15.36	$< .001$	0.54	0.39	0.64	0.66	1.92
	awa2	1.13	0.13	8.53	$< .001$	0.81	0.04	18.45	$< .001$	0.66				
	awa3	0.90	0.14	6.64	$< .001$	0.44	0.05	8.83	$< .001$	0.20				
collect	coll1	1.00^+				0.81	0.02	38.77	$< .001$	0.66	0.72	0.91	0.91	10.13
	coll2	0.76	0.05	14.86	$< .001$	0.76	0.04	20.99	$< .001$	0.58				
	coll3	1.06	0.04	23.63	$< .001$	0.94	0.01	70.89	$< .001$	0.88				
	coll4	0.95	0.05	18.14	$< .001$	0.86	0.03	33.94	$< .001$	0.74				
iuipc	collect	0.41	0.08	5.42	$< .001$	0.37	0.07	5.57	$< .001$	0.14				
	ctrl	0.42	0.07	6.05	$< .001$	0.61	0.09	6.47	$< .001$	0.38				
	aware	0.36	0.06	6.40	$< .001$	0.89	0.11	8.06	$< .001$	0.79				

Note: $^+$ fixed parameter; the standardized solution is STDALL

7.5 Respecification

In face of the disconfirming evidence discovered on construct validity and reliability, we decided to remove the items ctrl3 and awa3 from the scale, at the risk of losing identification. We stress that this step is not rooted in seeking a better fit, cautioned against a questionable specification practice [17, p. 641], but born from the disadvantagous properties of the items discussed. We compared the non-nesteds models IUIPC-10 and IUIPC-8 with the Vuong test on the ML estimation. The variance test indicated the two models as distinguishable, $\omega^2 = 1.926$, $p < .001$. The Vuong non-nested likelihood-ratio test rejected the null hypothesis that both models were equal. The IUIPC-8 model fitted statistically significantly better than the IUIPC-10 model, LRT $z = -34.541$, $p < .001$. Table 7 illustrates the comparison of the two models. This consitutes evidence of the factorial validity of the revised scale, including a confirmation of the unidimensionality of its sub-scales.

The respecification is still well correlated with IUIPC-10: (i) ctrl, $r = .91$, 95% CI $[.89, .93]$; (ii) awa, $r = .86$, 95% CI $[.83, .88]$; (iii) iuipc, $r = .96$, 95% CI $[.96, .97]$, all statistically significant at $p < .001$, yielding evidence for IUIPC-8's concurrent validity.

7.6 Validation

We validated the respecified IUIPC-8 model on the independent validation sample V and compared against the performance of IUIPC-10. We offered this comparison under consideration of all three estimators in Table 16.

First, we observed under ML estimation that the two models are indeed statistically significantly distinguishable with the variance test, $\omega^2 = 2.489$, $p < .001$. Furthermore, applied to the validation sample V, IUIPC-8 was the statistically significantly better model according to the Vuong test, LRT $z = -34.541, p < .001$.

For IUIPC-10 on V, we noticed that the ML estimation failed all test criteria, incl. the poor-fit hypothesis with a $\hat{\epsilon} = .08$ and $\hat{\epsilon}_{\mathsf{UL}} \geq .10$. The residuals showed a similar pattern as on Sample B. Again, the direct replication of IUIPC with outliers showed a poorer fit than the other ML estimations: $\chi^2(32) = 137.368, p < .001$; CFI=.93; RMSEA=.09 $[.07, .10]$; SRMR=.09; CAIC=270.5.

The respecified IUIPC-8 fared better. The ML estimator already offered a good fit. The estimators, MLM and WLSMVS, performed equally well, if not better.

Overall, we concluded that IUIPC-8 could be validated on an independent dataset as a well fitting model, whereas IUIPC-10 was disconfirmed once more.

Table 7. Respecification of IUIPC-10 to IUIPC-8 on Sample B.
All IUIPC-10 models failed the poor-fit tests irrespective of estimator. The respecified IUIPC-8 yielded a statistically significantly better fit by the Vuong test on the ML estimation, with better CAIC throughout. The RMSEA on IUIPC-8 still asks us to mind the residuals.

Instrument	Respecification	Estimator		
		ML	MLM‡	WLSMVS‡
IUIPC-10		$\chi^2(32) = 163.691, p < .001$ CFI=.92; GFI=1.00 RMSEA=.11 [.09, .12] SRMR=.10; CAIC=294.3	$\chi^2(32) = 131.417, p < .001$ CFI‡=.92; GFI=1.00 RMSEA‡=.10 [.08, .12] SRMR=.10; CAIC‡=262	$\chi^2(15) = 181.149, p < .001$ CFI‡=.95; GFI=.99 RMSEA‡=.17 [.15, .19] SRMR=.10; CAIC‡=414.6
	Trim ctrl3 & awa3	↑ ΔCFI $= 0.05$; ΔCAIC $= -132.57$ Vuong LRT $z = -34.541$, $p < .001$ ↓	↑ ΔCFI $= 0.05$; ΔCAIC $= -112.32$ ↓	↑ ΔCFI $= 0.04$; ΔCAIC $= -191.39$ ↓
IUIPC-8		$\chi^2(17) = 54.863, p < .001$ CFI=.97; GFI=1.00 RMSEA=.08 [.06, .10] SRMR=.03; CAIC=161.7	$\chi^2(17) = 42.836, p < .001$ CFI‡=.98; GFI=1.00 RMSEA‡=.07 [.05, .10] SRMR=.03; CAIC‡=149.7	$\chi^2(10) = 33.282, p < .001$ CFI‡=.99; GFI=1.00 RMSEA‡=.08 [.04, .12] SRMR=.04; CAIC‡=223.2

Note: ‡ Robust estimation with scaled test statistic. RMSEA reported wih 90% CI.

7.7 Summary

The construct validity of IUIPC according to RQ 2 bears a deliberate discussion. On the one hand, we could confirm the three-dimensionality and second-order model that Malhotra et al. [30] postulated. On the other hand, factorial validity evidenced in global and local fit—especially the unidimensionality of the first-order factors Control and Awareness—does not seem to be given, neither on the main Sample B nor on the validation Sample V. The convergent validity and reliability of these subscales is equally in question, our estimates having been lower than Malhotra et al.'s.

7.8 Estimator Appraisal

While we have already shown estimator invariance according to RQ 3 throughout confirmation and validation, we are now turning to the question of estimator viability asked in RQ 4.

Considering the viability of each estimator in their own right, ML without outlier treatment showed the worst performance of the ML-based estimations. We would discourage its use. Let us consider the remaining estimations with outlier treatment. Comparing between ML and MLM, both estimators behaved similarly on the relative changes between the misspecified IUIPC-10 and the respecified IUIPC-8. MLM consistently offered the stronger fit.

Even though robust WLS is estimating more parameters (loadings, errors, thresholds) than its ML counterparts, we still believe it is fair to say that WLSMVS seemed most affected by deviations from a close fit. On both samples, we observed a great improvement of fit indices when comparing between the WLSMVS IUIPC-10 and IUIPC-8 models. Hence, the robust WLS estimation is certainly viable in its own right, in fact, we would consider it quite robust against Type I errors.

Having assessed MLM and WLSMVS as viable on their global fit, let us further appraise their model interpretations and mean structures. To that end, the extended version of this paper [16] contains the complete threshold tables of the robust WLS estimation. For this inquiry, we follow the reasoning of Bovaird and Koziol [5, pp]. Let us train our lens on a single indicator for illustration: awa2. Here, the MLM estimation tells us that individuals with average levels of Awareness will have an expected response on item awa2 of 6.62. Of course, this value does not exist on the 7-point Likert scale. For each increase of one unit of Awareness, we would expect awa2 to increase by 1.13 points. For a single unit of increase from the mean, the prediction 7.75 is obviously out-of-bounds of the scale, hence, an invalid prediction. This phenomenon is an artefact of ML-based estimation on heavily non-normal, ordinal data.

Following the same line of inquiry for the robust WLS estimation, these issues do not exist. Here, the loading of awa2 $\lambda_{\text{awa2}} = 1.07$ indicates the expected change of the probit of an individuals's response to awa2 for each unit of Awareness. Under this estimation, an Awareness of -1.88 is required to choose response option 6–"Agree" or higher with a 50% probability; of -0.4 for option 7–"Strongly Agree" with a 50% probability.

Clearly, the robust WLS prediction of the individual's choice yields a more informative model.

8 Discussion

8.1 IUIPC-10 Could Not Be Confirmed

While we could attest to strengths in the underpinning of its content validity and the three-dimensionality of IUIPC-10, our CFA revealed a range of weaknesses in IUIPC's construct validity (cf. Table 8): (i) we could not confirm factorial validity in terms of global fit, meaning that the corresponding models did not approximate the corresponding observations well; (ii) the inspection of the residuals showed an unsatisfactory local fit for two items of the sub-scales of Control and Awareness, calling the the unidimensionality of these sub-scales into question. (iii) we found disconfirming evidence on the convergent validity. We further observed a sub-par reliability of the sub-scales Control and Awareness . As these assessments held true on main and validation samples, irrespective of estimators used, they offer disconfirming evidence against the scale itself.

For privacy researchers, the issues in construct validity mean that the IUIPC-10 scale shows weaknesses in measuring its hidden construct Information Privacy Concern. The observed sub-par reliability entails that IUIPC-10 measurements contain less common factor variance, which entails a low signal-to-noise ratio and can lead to spurious and erratic results.

8.2 IUIPC-8 Asserts a Stronger Validity

With IUIPC-8, we proposed a refined version of IUIPC that performed consistently well in terms of construct validity and reliability. We give an overview of a range of criteria in Table 8. In terms of factorial validity, we observed good global and local fits. Criteria for convergent and discriminant validity were fulfilled consistently. The respecified scale also showed appreciable improvements in reliability, yielding a 33% to 82% better signal-to-noise ratio for Awareness and Control, respectively.

We encourage privacy researchers consider carefully: On the one hand, the 10-item version of IUIPC exhibits a wider theoretical domain and contains more information on privacy concern they will care about. It has at least three items per factor and, thereby, creates favorable conditions for CFA model identification and more robust estimation of the true value of the latent factors. However, given that two items seem to misload to a considerable degree and to yield low reliability, those items may confound the model. The tight fit we obtained for the 8-item version of IUIPC is encouraging: it will approximate the data well and yield sound measurement models for subsequent analyses. The good concurrent validity with IUIPC-10 further supports using the respecified scale. Given the evidence in this study, we endorse adopting IUIPC-8 as a brief questionnaire for Internet users' privacy concern.

8.3 Questionnaire Wording as Culprit

While our reviews of IUIPC in Section 6 asserted sound content validity foundations, we equally found evidence for biases rooted in the questionnaire wording. Our analysis of the observed sample distributions of IUIPC-10 from Section 7.2 and especially the distribution graphs in Figure 3 showed a substantial negative skew and positive kurtosis. This seems to confirm our observation that the use of loaded words incl. "privacy" and "autonomy" may create a systematic bias through priming, further aggravated by leading questions. All these observations point towards the instrument itself influencing the respondents towards agreement.

IUIPC also suffers from instances of question wording yielding nondifferentiation. While the entire questionnaire can be subject to straightlining due to the absence of reverse-coded items, we believe that items ctrl3 and awa3 were especially impacted because of the presence of double-barreled constructions. This observation could explain why these items yielded a low reliability and why they needed to be removed altogether.

For privacy researchers, these observations stress the importance of inspecting the question wording of instruments carefully and to assess them against commonly known biases [8, 33].

8.4 How to Use IUIPC

Our analysis has a range of implications for privacy researchers, not just for the use of IUIPC but any multidimensional privacy concern scale. Let us consider assessing IUIPC scores for a given sample. Given the evidence that the instrument seems to bias responses towards agreement, it is invalid to call a particular sample "especially privacy-sensitive," should the mean IUIPC score be greater than 4–"Neither agree nor disagree."

Table 8. Selected evidence for construct validity and reliability criteria on Samples B and V under MLM estimation.
The factorial validity shows IUIPC-10 consistently failing fit measures; IUIPC-8 fared better, especially on the validation sample V.
The convergent validity of IUIPC-10 is evidenced to be consistently flawed; IUIPC-8 showed suitable results, where some standardized loadings β were shown to be border-line without violating $AVE > .50$. Divergent validity of IUIPC-10 suffers from the low AVE; IUIPC-8 fulfills all requirements. IUIPC-10 fails the requirements on internal consistency, especially $\omega > .70$; IUIPC-8 passes them.

		Construct Validity							Reliability	
		Factorial			Convergent		Divergent		Internal Consistency	
		$H_{\chi^2,0}$	$H_{\hat{\varepsilon}\leq.05,0}$	HBR	$\beta > .70$	$AVE > .50$	$\sqrt{AVE} > \forall\bar{r}$	$HTMT < .85$	$\alpha > .70$	$\omega > .70$
IUIPC-10	B	○	○	○	○	○	◐	●	○	○
	V	○	○	○	○	○	○	●	○	○
IUIPC-8	B	○	○	○	◐	●	●	●	●	●
	V	●	●	●	◐	●	●	●	●	●

Note: HBR = Hu and Bentler combination rule used by Malhotra et al. [30]; β = standardized loading; AVE = Average Variance Extracted; $\bar{r}$ = correlation with other factor; $HTMT$ = Heterotrait-Monotrait Ratio; ω = Composite Reliability.

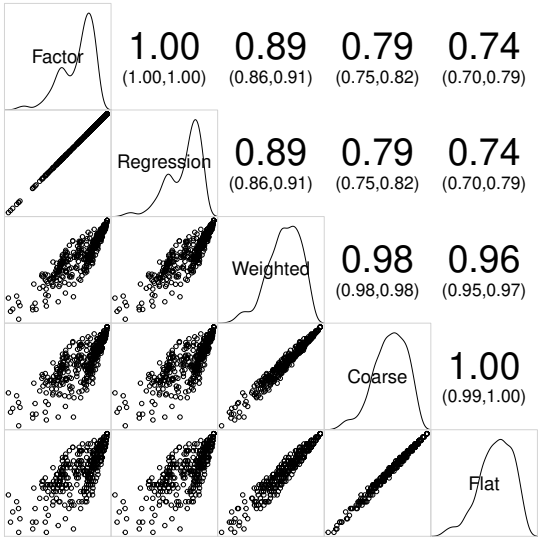


Fig. 4. Relation of the IUIPC-10 factors to approximate scores. While allowing for factor-analysis models also yielding estimates, overall, naïve score computations, such as flat averaging, lose information over more sophisticated factor scores.

Because IUIPC-10 exhibits a lower than desired reliability, privacy researchers need to be conscious of the low signal-to-noise ratio. The considerable uniqueness (specific and error variance of items) may mask information about the respondents' true IUIPC score.

The most important consequence of the low reliability for privacy researchers will be the expected attenuation of relations to other latent factors [36]: the magnitude of effects of IUIPC-measured privacy concern on other variables, say behavior, will be reduced. This means that it will be more difficult to show the impact of privacy concern—even if the true relation between the latent variables is substantive.

Privacy researchers focusing on simple statistical tools not factoring out error variance, such as linear regressions on summarized sub-scale scores, are most affected by these shortcomings: the specific and error variance is folded into the scores they use, masking the signal. For these researchers, IUIPC-8 offers considerable advantages by offering stronger validity and reliability. It comes at the price of eliminating two concern-items the researchers might be interested in. Let us consider the comparison of score approximations with the CFA-estimated factor in Figure 4 and Hair et al.'s introduction to summarized scales [17, pp. 160]. We assert that privacy researchers would be worst off ignoring the factor structure altogether by simply summing/averaging all items of IUIPC into one "flat" summarized score ($r = .74$). It is usually better to take into account the factor structure, the simplest approach averaging the sub-scale scores into a "coarse" summarized score. They could further improve their approximation by computing *factor scores*, for instance "weighted" by factor loadings ($r = .89$) or linearly combined with CFA-derived "regression" coefficients. While there is a great deal of discussion on correct and incorrect uses of those factor scores, a common pragmatic approach computes a weighting by the factor loading. However, this approach comes at a price of largely applying to the original sample and of offering less transferability. Given IUIPC-8's better factorial validity and reliability, coarse summarized score will be more generalizable.

Privacy researchers using advanced tools that exclude error variances, such as confirmatory factor analysis, face a different trade-off. The analysis will esti-

mate the loading of each indicator on the corresponding latent factor as well as the error variance. Hence, the separated-out error variance does not contaminate regression equations. However, the problematic construct validity of IUIPC-10 will affect their studies through appreciably weak global and local fit. They might discover their measurement model misspecified. For them, IUIPC-8 offers a better construct validity, especially useful for investigating comprehensive latent variable models in investigations of the privacy paradox.

In terms of choice of estimators for CFA and SEM (cf. Section 7.8), privacy researchers can viably opt for robust WLS as an estimator tailored to the non-normal, ordinal distributions of IUIPC indicators. As we have shown, a robust WLS estimation on IUIPC-8 offers a good fit at decent sample sizes. While the privacy community might not be accustomed to the interpretation of these probit models, they offer more nuanced model interpretability than the ML or MLM estimations. If privacy researchers decide against robust WLS, we would still advocate using a robust ML estimation (such as MLM or MLR) after a careful data preparation.

8.5 Limitations

In terms of generalizability, this study encountered the usual limitations of self-report instruments, where we sought to ward against straightlining with instructional manipulation checks/attention checks. While the sampling aimed at representing a UK population, the sampling process was non-randomized and affected by a self-selection bias due to Prolific's matchmaking.

Factor analyses like ours are affected by sampling and measurement errors. To ensure our findings yielded cross-sample validity, we considered three independent samples. To ensure our findings were valid irrespective of design decisions, we created a direct replication of the IUIPC-10 as well as analyses following methodological recommendations [24, p. 122]. The results were invariant to outlier inclusion and estimation method.

9 Conclusion

We independently evaluated the validity and reliability of IUIPC-10 in covariance-based confirmatory factor analyses. Acknowledging the sound content validity foundations presented by Malhotra et al. [30] and the instrumental role the scale played in advancing the field, we observed that (i) we could confirm the three-dimensionality of IUIPC-10 with the factors Control, Awareness, and Collection, (ii) we found disconfirming evidence for the factorial validity and convergent validity, largely rooted in evidence against the unidimensionality of Control and Awareness. These results were consistent in main analysis and validation, irrespective of the estimator used, (iii) we proposed a respecified IUIPC-8 that outperformed the original IUIPC-10 on construct validity and reliability consistently. (iv) we offered empirically grounded recommendations how to use IUIPC and similar privacy concern scales.

Future work specifically for IUIPC would ideally offer further carefully evaluated revisions to the scale, eliminating identified problems in question wording, e.g., eliminating loaded words, aiming for four items per factor, and establishing unassailable construct validity and an internal consistency $\omega > .80$ for all sub-scales.

We started this paper referring to instruments as measuring stick for privacy concern. If such a measurement stick is warped, the community's investigation of human factors of PETs and of the privacy paradox may be severely undermined and misled. We believe that we would benefit from concerted efforts to diligently evaluate standard instruments in the field along similar lines we have pursued in this study. While reaching consensus on sound measurement instruments on the construct Information Privacy Concern and its siblings is essential, we would also benefit from following unified recommendations their use.

Acknowledgments

We would like to thank Naresh K. Malhotra, James Agarwal and especially Sung S. Kim for promptly responding with succinct information on the original research process at the conception of IUIPC. We are grateful for the insightful comments of the anonymous reviewers of the Privacy-Enhancing Technology Symposium and especially for the guidance and time of our shepherd Michelle Mazurek. This work was supported by ERC Starting Grant CASCAde (GA n°716980).

References

[1] M. Ab Hamid, W. Sami, and M. Sidek. Discriminant validity assessment: Use of Fornell & Larcker criterion versus HTMT criterion. In Journal of Physics: Conference Series, volume 890, page 012163. IOP Publishing, 2017.

[2] J. C. Anderson, D. W. Gerbing, and J. E. Hunter. On the assessment of unidimensional measurement: Internal and external consistency, and overall consistency criteria. Journal of marketing research, 24(4):432–437, 1987.

[3] N. M. Barbosa, J. S. Park, Y. Yao, and Y. Wang. "what if?" predicting individual users' smart home privacy preferences and their changes. Proceedings on Privacy Enhancing Technologies, 2019(4):211–231, 2019.

[4] Z. Benenson, A. Girard, and I. Krontiris. User acceptance factors for anonymous credentials: An empirical investigation. In WEIS, 2015.

[5] J. A. Bovaird and N. A. Koziol. Measurement models for ordered-categorical indicators. In R. H. Hoyle, editor, Handbook of Structural Equation Modeling, pages 495–511. The Guilford Press, 2012.

[6] A. Braunstein, L. Granka, and J. Staddon. Indirect content privacy surveys: measuring privacy without asking about it. In Proceedings of the Seventh Symposium on Usable Privacy and Security, pages 1–14, 2011.

[7] T. Buchanan, C. Paine, A. N. Joinson, and U.-D. Reips. Development of measures of online privacy concern and protection for use on the internet. Journal of the American society for information science and technology, 58(2):157–165, 2007.

[8] B. C. Choi and A. W. Pak. Peer reviewed: a catalog of biases in questionnaires. Preventing chronic disease, 2(1), 2005.

[9] T. Dinev and P. Hart. Internet privacy concerns and their antecedents-measurement validity and a regression model. Behaviour & Information Technology, 23(6):413–422, 2004.

[10] C. DiStefano. The impact of categorization with confirmatory factor analysis. Structural equation modeling, 9(3):327–346, 2002.

[11] S. J. Finney and C. DiStefano. Non-normal and categorical data in structural equation modeling. Structural equation modeling: A second course, 10(6):269–314, 2006.

[12] C. Fornell and D. F. Larcker. Evaluating structural equation models with unobservable variables and measurement error. Journal of marketing research, 18(1):39–50, 1981.

[13] N. Gerber, P. Gerber, and M. Volkamer. Explaining the privacy paradox: A systematic review of literature investigating privacy attitude and behavior. Computers & Security, 77:226–261, 2018.

[14] N. Gerber, B. Reinheimer, and M. Volkamer. Investigating people's privacy risk perception. Proceedings on Privacy Enhancing Technologies, 2019(3):267–288, 2019.

[15] D. W. Gerbing and J. C. Anderson. An updated paradigm for scale development incorporating unidimensionality and its assessment. Journal of marketing research, 25(2):186–192, 1988.

[16] T. Groß. Validity and reliability of the scale internet users' information privacy concern (iuipc) [extended version]. arXiv.cs.HC Report arXiv:2011.11749, arXiv.org, Ithaca, New York, USA, 2020.

[17] J. F. Hair, W. C. Black, B. J. Babin, and R. E. Anderson. Multivariate data analysis. Cengage Learning, 8th edition, 2019.

[18] D. Harborth and S. Pape. German translation of the concerns for information privacy (cfip) construct. SSRN 3112207, 2018.

[19] D. Harborth and S. Pape. Jondonym users' information privacy concerns. In IFIP International Conference on ICT Systems Security and Privacy Protection, pages 170–184. Springer, 2018.

[20] D. Harborth and S. Pape. How privacy concerns and trust and risk beliefs influence users' intentions to use privacy-enhancing technologies-the case of tor. In Proceedings of the 52nd Hawaii International Conference on System Sciences, 2019.

[21] D. Harborth, S. Pape, and K. Rannenberg. Explaining the technology use behavior of privacy-enhancing technologies: The case of tor and jondonym. Proceedings on Privacy Enhancing Technologies, 2020(2):111–128, 2020.

[22] J. Henseler, C. M. Ringle, and M. Sarstedt. A new criterion for assessing discriminant validity in variance-based structural equation modeling. Journal of the academy of marketing science, 43(1):115–135, 2015.

[23] S. S. Kim. Data on context/reproducibility of iuipc. Personal Communication, May 2020.

[24] R. B. Kline. Assumptions in structural equation modeling. In R. H. Hoyle, editor, Handbook of Structural Equation Modeling, pages 111–125. The Guilford Press, 2012.

[25] R. B. Kline. Principles and practice of structural equation modeling. The Guilford Press, 4th ed. edition, 2015.

[26] S. Kokolakis. Privacy attitudes and privacy behaviour: A review of current research on the privacy paradox phenomenon. Computers & security, 64:122–134, 2017.

[27] C. E. Lance, M. M. Butts, and L. C. Michels. The sources of four commonly reported cutoff criteria: What did they really say? Organizational research methods, 9(2):202–220, 2006.

[28] P.-W. Lei and Q. Wu. Estimation in structural equation modeling. In R. H. Hoyle, editor, Handbook of structural equation modeling, pages 164–179. The Guilford Press, 2012.

[29] C.-H. Li. Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. Behavior Research Methods, 48(3):936–949, 2016.

[30] N. K. Malhotra, S. S. Kim, and J. Agarwal. Internet users' information privacy concerns (IUIPC): The construct, the scale, and a causal model. Information systems research, 15(4):336–355, 2004.

[31] S. Messick. Validity. ETS Research Report Series, 1987(2):i–208, 1987.

[32] A. Morton. Measuring inherent privacy concern and desire for privacy-a pilot survey study of an instrument to measure dispositional privacy concern. In 2013 International Conference on Social Computing, pages 468–477. IEEE, 2013.

[33] A. N. Oppenheim. Questionnaire design, interviewing and attitude measurement. Continuum, 1992.

[34] S. Preibusch. Guide to measuring privacy concern: Review of survey and observational instruments. International Journal of Human-Computer Studies, 71(12):1133–1143, 2013.

[35] Y. Pu and J. Grossklags. Towards a model on the factors influencing social app users' valuation of interdependent privacy. Proceedings on privacy enhancing technologies, 2016(2):61–81, 2016.

[36] W. Revelle and D. M. Condon. Reliability. In The Wiley Handbook of Psychometric Testing: A Multidisciplinary Reference on Survey, Scale and Test Development, pages 709–749. Wiley, 1st edition, 2018.

[37] K. S. Schwaig, A. H. Segars, V. Grover, and K. D. Fiedler. A model of consumers' perceptions of the invasion of information privacy. Information & Management, 50(1):1–12, 2013.

[38] J. C. Sipior, B. T. Ward, and R. Connolly. Empirically assessing the continued applicability of the iuipc construct. Journal of Enterprise Information Management, 2013.

[39] H. J. Smith, T. Dinev, and H. Xu. Information privacy research: an interdisciplinary review. MIS quarterly, pages 989–1015, 2011.

[40] H. J. Smith, S. J. Milberg, and S. J. Burke. Information privacy: measuring individuals' concerns about organizational practices. MIS quarterly, pages 167–196, 1996.

[41] K. A. Stewart and A. H. Segars. An empirical examination of the concern for information privacy instrument. Information systems research, 13(1):36–49, 2002.

[42] Q. H. Vuong. Likelihood ratio tests for model selection and non-nested hypotheses. Econometrica: Journal of the Econometric Society, pages 307–333, 1989.

[43] H. Xu, T. Dinev, H. J. Smith, and P. Hart. Examining the formation of individual's privacy concerns: Toward an integrative view. ICIS 2008 proceedings, page 6, 2008.

A Materials & Sample

We included the used IUIPC-10 questionnaire in Table 9. The questionnaire was administered in the first section of a greater survey, which included six instructional manipulation checks (IMCs) as attention checks shown in Table 10.

For the reproducibility of the maximum likelihood estimation, Table 11 contains the correlations and standard deviations (SDs) of Sample B. The OSF supplementary materials contain more precise covariance matrices of all samples.

B Additional Evidence

With respect to factorial validity in Section 7.3.1, Table 12 shows the comparison of candidate factor solutions. The path model from Figure 5 shows the selected IUIPC-10 model. Table 13 highlights the residuals founding the assessment of that model's local fit. Table 14 then offers the loadings and reliability of the respecified model from Section 7.5. Finally, Table 16 offers the fit comparison for the validation. The extended version of this paper [16] contains further tables on all aspects of the analysis.

C Factor Analysis

Factor analysis is a powerful tool for evaluating the construct validity and reliability of privacy concern instruments. *Factor analysis* refers to a set of statistical methods that are meant to determine the number and nature of *latent variables* (LVs) or *factors* that account for the variation and covariation among a set of observed measures commonly referred to as *indicators*.

In general, we distinguish *exploratory factor analysis* (EFA) and *confirmatory factor analysis* (CFA), both of which are used to establish and evaluate psychometric instruments, respectively. They are both based on the *common factor model*, which holds that each indicator variable contributes to the variance of one or more common factors and one unique factor. Thereby, *common variance* (communality) of related observed measures is attributed to the corresponding latent factor, and *unique variance* (uniqueness) seen either as variance associated with the indicator or as error variance. IUIPC is based on a *reflective measurement*, that is, the observed measure of an indicator variable is seen as caused by some latent factor. Indicators are thereby *endogenous* variables, latent variables *exogenous* variables. Reflective measurement requires that all items of the sub-scale are interchangeable [25, pp. 196].

In this paper, we are largely concerned with *covariance-based confirmatory factor analysis* (CB-CFA). Therein, the statistical tools aim at estimating coefficients for parameters of the measurement model that best fit the covariance matrix of the observed data. The difference between an observed covariance of the sample and an implied covariance of the model is called a *residual*.

C.1 Estimators and Their Assumptions

The purpose of a factor analysis is to estimate free parameters of the model (such as loadings or error variance), which is facilitated by *estimators*. The choice of

Table 9. Items of the instrument Internet users' information privacy concerns (IUIPC-10) [30]

Construct	Item	Question
Control (ctrl)	ctrl1	Consumer online privacy is really a matter of consumers' right to exercise control and autonomy over decisions about how their information is collected, used, and shared.
	ctrl2	Consumer control of personal information lies at the heart of consumer privacy.
	ctrl3	I believe that online privacy is invaded when control is lost or unwillingly reduced as a result of a marketing transaction.
Awareness (awa)	awa1	Companies seeking information online should disclose the way the data are collected, processed, and used.
	awa2	A good consumer online privacy policy should have a clear and conspicuous disclosure.
	awa3	It is very important to me that I am aware and knowledgeable about how my personal information will be used.
Collection (coll)	coll1	It usually bothers me when online companies ask me for personal information.
	coll2	When online companies ask me for personal information, I sometimes think twice before providing it.
	coll3	It bothers me to give personal information to so many online companies.
	coll4	I'm concerned that online companies are collecting too much personal information about me.

Note: The questionnaire is administered with 7-point Likert items, anchored on 1="Strongly Disagree" to 7="Strongly Agree"

Table 10. Items of our instructional manipulation checks

Item	Question
A1	It is important you pay attention to the statements. Please agree by choosing 'strongly agree'.
A2	To confirm that you are paying attention to the questions in the questionnaire, please select the first option from the left on the scale.
A3	I'm paying attention to the questions in this questionnaire. I confirm this by choosing 'somewhat agree'.
A4	I recognise the importance of paying attention to the questions in this questionnaire. Please select 'agree' to confirm your agreement.
A5	Paying attention to the questions in this questionnaire is important. I agree by choosing the third option from the left of the scale.
A6	When you're responding to the questions in the questionnaire it is important that you're paying attention. Please agree by selecting the second option from the left on the scale.

estimator matters, because each comes with different strengths and weaknesses, requirements and assumptions that need to be fulfilled for the validity of their use. The most commonly used method for confirmatory factor analysis is *maximum likelihood* (ML) estimation. Among other *assumptions*, this estimator requires according to Kline [25, pp. 71]: (i) a *continuous measurement level*, (ii) *multi-variate normal distribution* (entailing the absence of extreme *skewness*) [25, pp. 74], and (iii) treatment of influential cases and *outliers*. The distribution requirements are placed on the endogenous variables: the indicators.

These requirements are not always fulfilled in samples researchers are interested in. For instance, a common case at odds with ML-based CFA is the use of Likert items as indicator variables. Likert items are *ordinal* [17, p. 11] in nature, that is, ordered categories in which the distance between categories is not constant; they thereby require special treatment [25, pp. 323].

Lei and Wu [28] held based on a number of empirical studies that the fit indices of approximately normal ordinal variables with at least five categories are not greatly misleading. However, when ordinal and non-normal is treated as continuous and normal, the fit is underestimated and there is a more pronounced negative bias in estimates and standard errors. While Bovaird and Kozoil [5] acknowledge robustness of the ML estimator with normally distributed ordinal data, they stress that increasingly skewed and kurtotic ordinal data inflate the Type I error rate. In the same vein, Kline [24, p. 122] holds the normality assumption for endogenous variables—the indicators—to be critical.

C.2 Global and Local Fit

The *closeness of fit* of a factor model to an observed sample is evaluated globally with fit indices as well as locally by inspecting the residuals. We shall focus on the ones Kline [25, p. 269] required as minimal reporting.

$\chi^2(df)$: The χ^2 for given degrees of freedom is the likelihood ratio chi-square, as a measure of exact fit.

Table 11. Correlations and Standard Deviations of Sample B

	1	2	3	4	5	6	7	8	9	10
1. ctrl1										
2. ctrl2	0.56									
3. ctrl3	0.25	0.27								
4. awa1	0.25	0.23	0.25							
5. awa2	0.32	0.32	0.30	0.62						
6. awa3	0.23	0.19	0.26	0.32	0.31					
7. coll1	0.05	0.05	0.26	0.05	0.11	0.33				
8. coll2	0.10	0.06	0.28	0.22	0.24	0.30	0.66			
9. coll3	0.16	0.09	0.32	0.16	0.22	0.40	0.76	0.72		
10. coll4	0.19	0.10	0.31	0.23	0.23	0.47	0.71	0.62	0.81	
SD	0.93	0.93	1.00	0.55	0.56	0.82	1.36	1.11	1.24	1.22

Note: $N_B = 370$

Table 12. Comparison of different model structures of IUIPC-10 on Sample B with MLM estimation.
The models show increasingly better fits in scaled χ^2, CFI, and CAIC, supporting the predicted hierarchical three-factor model.

	One Factor		Two Factors		Three Factors (1st Order)		Three Factors (2nd Order)	
$\chi^2(df)$	481.87 (35)		239.28 (34)		163.69 (32)		163.69 (32)	
χ^2/df	13.77		7.04		5.12		5.12	
CFI	.73		.87		.92		.92	
GFI	1.00		1.00		1.00		1.00	
RMSEA	.19	[.17, .20]	.13	[.11, .14]	.11	[.09, .12]	.10	[.09, .12]
SRMR	.14		.09		.10		.10	
Scaled $\chi^2(df)$	377.87 (35)		189.71 (34)		131.42 (32)		131.42 (32)	
CAIC	600.572		361.941		294.264		294.264	
Scaled CAIC	496.570		312.366		261.990		261.990	

CFI: The *Bentler Comparative Fit* Index is an incremental fit index based on the non-centrality measure comparing selected against the null model.

RMSEA: *Root Mean Square Estimate of Approximation* ($\hat{\varepsilon}$) is an absolute index of bad fit, reported with its 90% Confidence Interval $[\hat{\varepsilon}_{\mathsf{UL}}, \hat{\varepsilon}_{\mathsf{LL}}]$.

SRMR: *Standardized Root Mean Square Residual* is a standardized version of the mean absolute covariance residual, where zero indicates excellent fit.

We mention the *Goodness-of-Fit index* (GFI) reported by IUIPC, which approximates the proportion of variance explained in relation to the estimated population covariance. It is not recommended as it is substantially impacted by sample size and number of indicators.

Malhotra et al. [30] adopted a *combination rule* referring to Hu and Bentler, which we will report as HBR, staying comparable with their analysis: "A model is considered to be satisfactory if (i) CFI $> .95$, (ii) GFI $> .90$, and (iii) RMSEA $< .06$."

Nested models [25, p. 280], that is, models with can be derived from each other by restricting free parameters, can be well-compared with a *Likelihood Ratio* χ^2 *Difference Test* (LRT) [25, p. 270]. However, the models we are interested in are *non-nested* [25, p. 287], because they differ in their observed variables. On ML-estimations, we have the *Vuong Likelihood Ratio Test* [42] at our disposal to establish statistical inferences on such models.

In addition, we introduce the AIC/BIC family of metrics with formulas proposed by Kline [25, p. 287]:

$$\mathsf{AIC} := \chi^2 + 2q \qquad \mathsf{BIC} := \chi^2 + q\ln(N),$$

where q is the number of free parameters and N the sample size. We compute CAIC as the even-weighted mean between AIC and BIC. These criteria can be used to compare different models estimated on the same samples, on the same variables, but theoretically also on different subsets of variables. Smaller is better.

Statistical Inference

The χ^2 and RMSEA indices offer us *statistical inferences of global fit*. Such tests can either be *accept-support*, that is, accepting the null hypothesis supports the selected

Table 13. Residuals of the MLM-estimated CFA of IUIPC-10 on Sample B.
The highlighted distinctive residual patterns show a poor local fit. They indicate misloadings of ctrl3 and awa3.

(a) Correlation residuals

	1	2	3	4	5	6	7	8	9	10
1. ctrl1	—									
2. ctrl2	0.021	—								
3. ctrl3	-0.046	-0.022	—							
4. awa1	-0.042	-0.062	0.089	—						
5. awa2	-0.01	0	**0.115**	0.018	—					
6. awa3	0.048	0.015	**0.163**	-0.007	-0.049	—				
7. coll1	-0.082	-0.085	**0.189**	**-0.147**	**-0.111**	**0.206**	—			
8. coll2	-0.025	-0.068	**0.211**	0.034	0.032	**0.186**	0.037	—		
9. coll3	0.001	-0.069	**0.234**	-0.065	-0.031	**0.264**	-0.007	0.003	—	
10. coll4	0.051	-0.043	**0.231**	0.024	0.003	**0.345**	0.006	-0.038	0.004	—

Note: Correlation residuals in absolute > 0.1 are marked

(b) Standardized residuals

	1	2	3	4	5	6	7	8	9	10
1. ctrl1	—									
2. ctrl2	**3.703**	—								
3. ctrl3	**-2.345**	-1.016	—							
4. awa1	-1.585	**-2.466**	1.921	—						
5. awa2	-0.457	0.003	**2.597**	**4.618**	—					
6. awa3	0.972	0.379	**3.012**	-0.36	**-3.891**	—				
7. coll1	**-2.304**	**-2.617**	**3.324**	**-4.525**	**-3.544**	**3.878**	—			
8. coll2	-0.669	-1.843	**4.334**	0.706	0.845	**3.971**	**2.035**	—		
9. coll3	0.031	**-2.609**	**4.142**	**-2.448**	-1.365	**5.144**	-1.384	0.405	—	
10. coll4	1.293	-1.429	**4.284**	0.688	0.101	**5.912**	0.475	**-2.636**	1.397	—

Note: Statistically significant residuals (abs > 1.96) are marked

Table 14. Factor loadings and their standardized solution of the MLM CFA of IUIPC-8 on Sample B.
The standardized factor loadings β are largely greater than .70, yielding satisfactory $AVE > .50$. The construct reliability $\omega > .70$ passes the threshold for control and awareness yielding moderate signal-to-noise ratios; the reliability of collection is excellent.

Factor	Indicator	Factor Loading				Standardized Solution				Reliability				
		λ	SE_λ	Z_λ	p_λ	β	SE_β	Z_β	p_β	R^2	AVE	α	ω	S/N_ω
ctrl	ctrl1	1.00^+				0.76	0.06	13.40	$< .001$	0.57	0.56	0.72	0.72	2.52
	ctrl2	0.98	0.14	7.01	$< .001$	0.74	0.06	11.58	$< .001$	0.54				
aware	awa1	1.00^+				0.69	0.05	13.45	$< .001$	0.48	0.64	0.76	0.78	3.55
	awa2	1.33	0.18	7.33	$< .001$	0.90	0.05	17.34	$< .001$	0.80				
collect	coll1	1.00^+				0.81	0.02	39.01	$< .001$	0.66	0.72	0.91	0.91	10.13
	coll2	0.76	0.05	14.83	$< .001$	0.76	0.04	20.90	$< .001$	0.58				
	coll3	1.06	0.04	23.67	$< .001$	0.94	0.01	70.72	$< .001$	0.88				
	coll4	0.95	0.05	18.15	$< .001$	0.86	0.03	33.67	$< .001$	0.74				
iuipc	collect	0.34	0.08	4.28	$< .001$	0.31	0.07	4.31	$< .001$	0.09				
	ctrl	0.39	0.08	4.90	$< .001$	0.56	0.11	5.01	$< .001$	0.31				
	aware	0.33	0.07	4.99	$< .001$	0.86	0.14	6.03	$< .001$	0.74				

Note: $^+$ fixed parameter; the standardized solution is STDALL

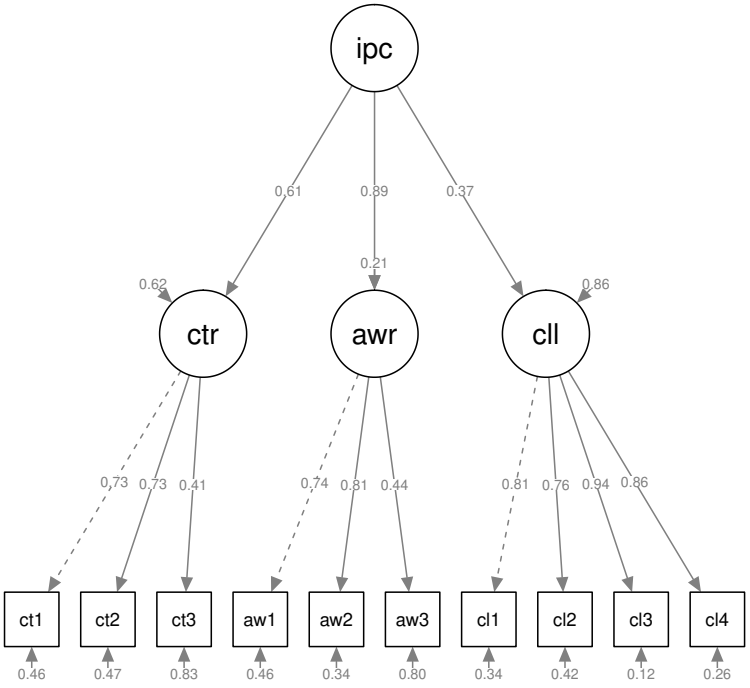


Fig. 5. CFA paths plot with standardized estimates of IUIPC-10 on Sample B. Note: The coefficients are standardized

Table 15. Evidence for discriminant validity of IUIPC-10 (Fornell-Larcker Criterion and Heterotrait-Monotrait Ratio) on Sample B. The Fornell-Larcker criterion of $\sqrt{AVE}$ greater than any correlation with any other factor and the $HTMT < .85$ are fulfilled.

(a) Fornell-Larcker

	1	2	3	4
1. ctrl	0.634			
2. aware	0.546	0.627		
3. collect	0.227	0.329	0.848	
4. iuipc	0.613	0.891	0.37	0.769

Note: The diagonal contains the $\sqrt{AVE}$

(b) HTMT Ratio

	1	2	3
1. ctrl	—		
2. aware	0.67	—	
3. collect	0.32	0.45	—

Table 16. Comparison of IUIPC-10 and IUIPC-8 on Validation Sample V.
The IUIPC-10 models fail the fit tests irrespective of estimator. By Vuong test on the ML estimation and the CAIC, the IUIPC-8 models are better fits than their corresponding IUIPC-10 equivalents. The IUIPC-8 models show very good fit, with the MLM estimation passing even the exact-fit test.

Instrument	Respecification	Estimator		
		ML	MLM‡	WLSMVS‡
IUIPC-10		$\chi^2(32) = 122.015, p < .001$ CFI=.94; GFI=1.00 RMSEA=.08 [.07, .10] SRMR=.08; CAIC=254.6	$\chi^2(32) = 80.69, p < .001$ CFI=.95; GFI=1.00 RMSEA=.07 [.05, .09] SRMR=.08; CAIC=213.3	$\chi^2(16) = 151.118, p < .001$ CFI=.96; GFI=.99 RMSEA=.14 [.12, .16] SRMR=.07; CAIC=404.3
	Trim ctrl3 & awa3	↑ ΔCFI $= 0.05$; ΔCAIC $= -111.6$ Vuong LRT $z = -35.146$, $p < .001$ ↓	↑ ΔCFI $= 0.05$; ΔCAIC $= -82.76$ ↓	↑ ΔCFI $= 0.04$; ΔCAIC $= -174.5$ ↓
IUIPC-8		$\chi^2(17) = 34.532, p = .007$ CFI=.99; GFI=1.00 RMSEA=.05 [.03, .07] SRMR=.03; CAIC=143	$\chi^2(17) = 22.04, p = .183$ CFI=.99; GFI=1.00 RMSEA=.03 [.00, .07] SRMR=.03; CAIC=130.6	$\chi^2(10) = 24.844, p = .005$ CFI=.99; GFI=1.00 RMSEA=.06 [.03, .09] SRMR=.03; CAIC=229.8

Note: ‡ Robust estimation with scaled test statistic.

model, or *reject-support*, that is, rejecting the null hypothesis supports the selected model. We present them in the order of decreasing demand for close approximation.

***Exact Fit*:** An accept-support test, in which rejecting the null hypothesis on the model χ^2 test implies the model is not an exact approximation of the data. The test is sensitive to the sample size and may reject well-fitting models at greater N.

$H_{\chi^2,0}$**:** The model is an exact fit in terms of residuals of the covariance structure.

$H_{\chi^2,1}$**:** The residuals of the model are considered too large for an exact fit.

***Close Fit*:** An accept-support test, evaluated on the RMSEA $\hat{\varepsilon}$ with zero as best result indicating approximate fit.

$H_{\varepsilon_0 \leq .05,0}$**:** The model has an approximate fit with RMSEA being less or equal .05.

$H_{\varepsilon_0 \leq .05,1}$**:** The model does not evidence a close fit.

***Not-close Fit*:** A reject-support hypothesis operating on the upper limit if the RMSEA 90% CI, $\hat{\varepsilon}_{\mathsf{UL}}$, a significant p-value rejecting the not-close fit.

$H_{\varepsilon_0 \geq .05,0}$**:** The model is not a close fit, $\hat{\varepsilon}_{\mathsf{UL}} \geq .05$.

$H_{\varepsilon_0 \geq .05,1}$**:** Model approximate fit, $\hat{\varepsilon}_{\mathsf{UL}} < .05$.

***Poor Fit*:** A reject-support test on the upper limit if the RMSEA 90% CI, $\hat{\varepsilon}_{\mathsf{UL}}$ checking for a poor fit.

$H_{\varepsilon_0 \geq .10,0}$**:** The model is a poor fit with $\hat{\varepsilon}_{\mathsf{UL}} \geq .10$.

$H_{\varepsilon_0 \geq .10,1}$**:** Model not a poor fit, $\hat{\varepsilon}_{\mathsf{UL}} < .10$.

Even with excellent global fit indices, the inspection of the local fit—evidenced by the residuals—must not be neglected. In fact, Kline [25, p. 269] drives home "Any report of the results without information about the residuals is incomplete." Simply put, a large absolute residual indicates covariation that the model does not approximate well and that may thereby lead to spurious results.