Tom Biselli, Enno Steinbrink, Franziska Herbert, Gina M. Schmidbauer-Wolf, and Christian Reuter

On the Challenges of Developing a Concise Questionnaire to Identify Privacy Personas

Abstract: Concise instruments to determine privacy personas – typical privacy-related user groups – are not available at present. Consequently, we aimed to identify them on a privacy knowledge–privacy behavior ratio based on a self-developed instrument. To achieve this, we conducted an item analysis (N = 820) and a confirmatory factor analysis (CFA) (N = 656) of data based on an online study with German participants. Starting with 81 items, we reduced those to an eleven-item questionnaire with the two scales privacy knowledge and privacy behavior. A subsequent cluster analysis (N = 656) revealed three distinct user groups: *(1) Fundamentalists* scoring high in privacy knowledge and behavior, *(2) Pragmatists* scoring average in privacy knowledge and behavior and *(3) Unconcerned* scoring low in privacy knowledge and behavior. In a closer inspection of the questionnaire, the CFAs supported the model with a close global fit based on RMSEA in a training and to a lesser extent in a cross-validation sample. Deficient local fit as well as validity and reliability coefficients well below generally accepted thresholds, however, revealed that the questionnaire in its current form cannot be considered a suitable measurement instrument for determining privacy personas. The results are discussed in terms of related persona conceptualizations, the importance of a methodologically sound investigation of corresponding privacy dimensions and our lessons learned.

Keywords: privacy, personas, clustering, questionnaire

DOI 10.2478/popets-2022-0126
Received 2022-02-28; revised 2022-06-15; accepted 2022-06-16.

1 Introduction

With an ever increasing part of our daily lives taking place online, each and every individual leaves behind an ever increasing amount of digital traces. This fuels an interest in such digital traces by both private corporations and state institutions. Negative effects of this "hunger for data" became clear, for example, in the infamous case of Cambridge Analytica, where data was analyzed and misused for political purposes. Actual practices therefore stood in stark contrast to limits of user consent and the perceived responsibilities and obligations of private corporations' management of user data [36]. Thus, individuals generally should have an interest in maintaining their privacy in an increasingly digitalized world. The interest in personal data privacy is confirmed by surveys, in which individuals consistently claim that privacy is important to them [1]. However, these attitudes do not necessarily coincide with online behavior, a phenomenon also known as the *privacy paradox* [1, 2]. This indicates two things: first, users seem to require some kind of assistance in order to to act in accordance with their own attitudes. Second, user attitude and behavior apparently have not been sufficiently understood until now. This is problematic, both from a user perspective and for developers trying to create products which satisfy user needs.

One reason for this lack of understanding of typical users could be the fact that there is no such thing as the typical user whose privacy requirements must be met. A "one-solution-fits-all" approach might not sufficiently account for the diversity of user demands. If tools are to be developed which assist users and consider their privacy demands, a more fine-grained approach seems promising. Such an individualization has previously been proposed as a suitable approach for supporting users in making better privacy decisions [43]. A central challenge in the context of individualization, however, is the lack of concise ways to determine individual user needs. While several approaches exist to cluster privacy personas, the development of a methodologically sound instrument remains challenging.

Thus, the objective of the present study was to identify privacy personas — specifications of typical user groups – and to provide insights into the methodological evaluation of a self-developed concise questionnaire. Although we obtained a factor model with acceptable global fit measures, the resulting inventory had a low

Tom Biselli, Enno Steinbrink, Franziska Herbert, Gina M. Schmidbauer-Wolf, Christian Reuter: Science and Technology for Peace and Security (PEASEC) Technical University of Darmstadt, E-mail: <lastname>@peasec.tu-darmstadt.de

validity and reliability. The questionnaire should therefore not be used in its current form. Nonetheless, in attempting to account for ordinality of items and to perform adequate cross-validation, among other things, we encountered several methodological challenges that we would like to share with the research community. The core contributions of this work lie in (1) creating awareness for the relevance of methodologically sound instruments to determine privacy personas, (2) transparent reporting when establishing questionnaires and (3) illuminating that a failure to apply such a thorough questionnaire analysis is one reason for the heterogeneity in the number of privacy personas found in the literature. The lessons learned in this process should make it easier for other researchers to identify privacy personas on the basis of methodologically sound instruments.

2 Related Work

Understanding users in software development is of utmost relevance. Especially concerning the development of user-centered privacy-enhancing technologies (PETs), a suitable segmentation of users based on relevant privacy dimensions is crucial for user acceptance.

2.1 The Myth of the Average User

During the software design process, it is tempting to target an average user. However, previous research has shown that the "average user" might actually be a myth and that users differ substantially in their privacy preferences and needs [4, 20, 25]. Accordingly, various individual differences exist, for example in terms of personality traits and decision-making styles in general, which should be considered in the privacy realm. It was proposed that users' adaptation to privacy risks consists of highly complex and multifaceted coping strategies, including problem-focused, emotion-focused, and communication-focused [14]. Such complex processes illustrate the large potential for individual variability in corresponding attitudes, demands, and behavior. Individualization would thus be the key to designing effective privacy and security architectures. Generally, while some users actively try to protect their privacy by using PETs, others care very little about their privacy [29]. Furthermore, differences regarding privacy concerns and perceptions not only exist on individual levels, but also between cultures [13]. Thus, a one-size-fits-all approach might be limited, which is why a clear understanding of different user types is vital.

Apart from individual differences in personality traits and decision-making styles, other factors have previously been assumed to be relevant with regard to privacy behavior and attitudes. For example, in terms of political ideology, studies suggest that people who see themselves as rather left-wing are more critical of (predominantly state-organized) data collection on individuals [1, 8, 63]. Further differences have been reported, especially if attitudes and behaviors concerning security are taken into account, as privacy and security are related to a certain extent [9]. For example, analyzing gender-differential aspects, studies demonstrate that women on average show less security knowledge, experience, and behavior than men [32, 47]. Taking the age of users into account, younger individuals (<25 years) have been associated with weaker security behavior [38, 68]. Due to the association between security and privacy [9], findings from the security domain might also be applicable to the privacy domain. In the privacy domain, older adults have been been associated with being more concerned about privacy [72] and disclosing less information on social media [40]. Also, a meta-analysis found that women display higher privacy concerns and behaviors on a social network site than men [71]. Finally, it has been shown that those with higher levels of education tend to be more concerned about privacy [51] and show more security awareness [52].

A more general framework for explaining determinants and differences in privacy and security behavior was provided with a a three-step model [18]. First, privacy is described as "economic rationality" and the result of a trade-off between cost and benefit concerning the disclosure of information. Second, privacy is described as "practical action" with a focus on in-the-moment actions without the need for abstract descriptions of broader influencing factors. Third, privacy is described as a "discursive practice", which determines the separation between secure and insecure actions. Depending on which model is applied to conceptualize privacy and security decisions, vastly different influential factors on decisions and behavior are identified.

Taken together, a complex picture emerges, showing that there are plenty of sources of variability between individuals. Especially considering that individualized user interventions have previously proven promising, both in the security [30] and privacy [43] domain, structuring the user space to better address individual user needs is a promising path to pursue.

2.2 User Typologies

Previous attempts to categorize users often relied on the segmentation index proposed by Westin, which categorized users in (1) *Fundamentalists*, (2) *Pragmatists*, and (3) *Unconcerned* [45]. Based on this categorization, several approaches have been undertaken to illuminate privacy demands of users. It has been shown, however, that these categories, behavioral intentions, and the consequences of privacy behaviors do not necesarrily correlate [74]. Generally, the accuracy and practical significance of Westin's three privacy categories have been challenged in empirical studies and by exposing methodological weaknesses of the instrument used to determine the privacy personas [15, 39, 41, 57]. This has led to rethinking and refining the original approach, including further developments. To illustrate differences in the aging population, studies highlight that with more elderly people engaging with the digital world, they might show specific online behavior. In this context, one study built on the the typology of Westin and distinguished five users types based on the findings from 40 interviews [21]. Here, older adults' (65+ years) attitudes toward privacy were found to vary widely, with a large proportion of older adults being only marginally concerned and considering the frequency of their online activities to be low and therefore not particularly relevant. As a result, targeted training for elderly users to enhance their privacy literacy was proposed.

Dupree et al., on the other hand, collected qualitative and quantitative data to cluster users into five categories based on privacy knowledge and privacy motivation [19]. Besides the known categories of *Fundamentalists* and *Marginally Concerned* (similar to *Unconcerned*), *Lazy Experts*, *Self-Educated Technicians*, and *Amateurs* were identified here to allow for a more nuanced picture. However, the psychometric properties of the survey items used and further information on scale validity and reliability were not provided. This categorization was adopted by another study, which showed that a better understanding of certain user groups can increase effectiveness and efficiency with regard to dealing with privacy settings for some users [62].

Other approaches have also led to a more fine-grained differentiation of the user space with five proposed user categories. For example, through a combined research design of qualitative and quantitative methods, users were segmented in five user groups according to the information cues considered important by users in the context of a technology service [50]. Instead of answering a questionnaire with responses on a Likert scale, here participants had to sort topics according to individually perceived importance. Based on the question, whether types of privacy concerns online are mirrored in the offline environment, it was also proposed that user concerns might be best represented via four groups [66]. An overview of previously proposed persona conceptualizations can be found in Table 1.

A general distinction of user segmentations lies in the specific domains of privacy on which they are based. While some studies focus on privacy motivation, others focus on privacy behavior or attitudes. Sometimes, a combination is used for a more comprehensive conceptualization. With a focus on attitudes, one study applied a self-developed privacy orientation scale measuring privacy attitudes of social media users [6]. The scale development was thorough, considered existing privacy scales and a pilot study was conducted. Additionally, exploratory factor analysis and confirmatory factor analyses were applied to evaluate the construct validity of the scale and reliability issues were reported. Based on the resulting questionnaire, the following three user groups were segmented: *Privacy Advocates*, *Privacy Individualists*, and *Privacy Indifferents* [6]. Another study set the focus on privacy behavior and examined posting behavior on social media. Here, users were segmented on the two dimensions "content appropriateness" and "privacy concern" [48]. Finally, one study combined the dimensions of user concerns and behavior by conducting qualitative interviews and a quantitative survey. On that basis users were segmented into the dimensions of "privacy concerns" and "privacy behavior" [65]. The questionnaire development here also included a brief report on reliability and (convergent and discriminant) validity indices - without a more thorough evaluation of the construct validity using factor analyses.

2.3 Measuring Privacy Knowledge and Behavior

There have been previous attempts to quantify privacy knowledge and behavior. On the topic of privacy concerns, the Internet Users Privacy Concerns (IUIPC) represents a well established instrument [56] as one of several other much cited instruments trying to quantify privacy concerns [17, 67]. One instrument applied to measure online privacy competency is the "Online Privacy Literacy Scale" (OPLIS) [46]. The OPLIS is fairly comprehensive and measures privacy literacy on four dimensions in relation to (1) institutional practices, (2) technical aspects of privacy, (3) legal aspects of privacy,

Authors	Method	Privacy Dimensions	Privacy Personas
Elueze et al. [21]	Qualitative Interviews	Qualitative multifactorial	(1) Fundamentalist, (2) Intense Pragmatist, (3) Relaxed Pragmatist, (4) Marginally Concerned
Dupree et al. [19]	Qualitative Interviews, Quantitative Survey	Knowledge & Motivation	(1) Fundamentalist, (2) Lazy Expert, (3) Technician, (4) Amateurs, (5) Marginally Concerned
Schomakers et al. [65]	Qualitative Interviews, Quantitative Survey	Attitude & Behavior	(1) Privacy Guardian, (2) Privacy Cynic, (3) Privacy Pragmatist
Morton et al. [50]	Q methodology	Organizational, Technology lens, Other factors	(1) Information Controllers, (2) Security Concerned, (3) Benefits Seekers, (4) Crowd Followers, (5) Organizational Assurance Seekers

Table 1. Previous Privacy Persona Conceptualizations.

and (4) privacy strategies. Due to its comprehensiveness this instrument is not suitable for a succinct assignment of privacy personas. Furthermore, the OPLIS is not necessarily universally applicable, since the subscale covering legal aspects of privacy is specific to European data regulation frameworks. Another instrument tries to combine cognitive and behavioral aspects by using three scales that measure privacy-related attitudes (Privacy Concern) and behaviors (General Caution and Technical Protection) [11]. Concerning privacy behavior, a generally established scale has not yet been developed. Sometimes, a behavioral dimension is contained within a multifactorial questionnaire, as previously described [11]. One reason for the absence of explicit scales for privacy-related behavior is that this is ideally measured by actually observing users' behavior. However, most studies still make use of self-reports as an approximation to actual behavior. Often, specific scenarios are presented in which individuals report the likelihood of engaging in a certain behavior, like regularly reading terms and conditions before installing potentially privacy-invasive apps. Sometimes, instruments focusing on privacy issues are intertwined with security issues. In the context of information security awareness, the Human Aspects of Information Security Questionnaire (HAIS-Q) represents a well established instrument. The HAIS-Q measures information security awareness with a focus on information security threats caused by employees within organizations [55]. For the validation of instruments such as those mentioned above, different approaches have been taken: Mainly a new confirmatory factor analysis model is fit for the same item measures to a new sample based on the factor structure to be validated (but usually with new parameter estimates, e.g. [53, 76]). Alternatively, an exploratory factor analysis is conducted to evaluate the underlying latent variables (e.g. [69]). Sometimes these approaches are combined [3]. Interestingly, even widely used instruments such as the IUIPC showed weaknesses in terms of its dimensionality, validity and reliability when thoroughly evaluated in terms of its psychometric properties [27].

Overall, some instruments on aspects of privacy and security already exist, sometimes well validated and sometimes less so. All such instruments have in common that they try to conceptualize privacy or information security constructs in a fairly comprehensive manner. However, none of the described instruments explicitly conceptualize privacy knowledge and privacy behavior as two separate dimensions within one questionnaire with a focus on conciseness.

2.4 Research Gap

Since users differ in their privacy needs, the categorization of users is a relevant research field. Several of the previously mentioned applications of privacy personas rely on Westin's Segmentation Index [45], which has sometimes failed to provide relevance for behavioral privacy scenarios. Further developments of categorizations of personas focus on different aspects of privacy, including concerns, motivation, literacy, or behavior. None of the existing typologies, however, solely use privacy knowledge and privacy behavior as two distinct dimensions in a quantitative survey with a focus on concise questionnaire development. The present segmentation is generally oriented at previous privacy segmentations. Dupree et al. [19], for example, showed that privacy personas can be clustered by combining multiple dimensions and qualitative and quantitative approaches, whereby the dimensions of privacy knowledge, motivation, and behavior were considered relevant. Our approach, on the other hand, specifically focuses on trying to establish a concise questionnaire for the two dimensions of privacy knowledge and behavior as the two dimensions for defining a user space. This focus is particularly promising since privacy knowledge and behavior

are strongly related, and knowledge about privacy issues exerts a significant influence on corresponding behavior. In this context, a "mobile privacy-security knowledge gap model" was proposed with the goal to provide a framework for explaining privacy and security behavior. This model emphasizes that besides information about privacy issues, motivation, and belief, knowledge is a decisive factor for explaining subsequent privacy behavior [16]. Furthermore, previous segmentations have typically focused on elaborate methods and sometimes combinations of qualitative and quantitative approaches to create highly sophisticated and differentiated descriptions of privacy personas. However, few studies focused on the development of a quantitative instrument to determine privacy personas in a time-saving manner with transparent reporting of the challenges arising when trying to establish a methodologically sound instrument.

2.5 Aims and Hypotheses

Based on the previously elaborated research gap, we wanted to address the following research question: **"Can a concise questionnaire be developed for defining a user space with distinct privacy personas using a knowledge-behavior ratio?"** The subordinate objectives associated with this research question included (1) establishing a suitable model of the questionnaire (H1 & H2) and (2) clustering users based on resulting privacy knowledge and behavior scores (H3). With previous studies having suggested that age, gender and education are relevant for shaping privacy attitudes and behaviors, another objective also included (3) analyzing characteristics of resulting privacy personas (H4). Our hypotheses thus included:

- **H1**: A factor analysis confirms the two-factor structure of the questionnaire (knowledge - behavior) with a close model fit.
- **H2**: The resulting model can be cross-validated in an independent subsample with a close fit.
- **H3**: A cluster analysis reveals different privacy personas.
- **H4**: Demographic factors (age, gender, education) can predict membership in a privacy persona group.

3 Method

In order to categorize privacy personas, item analysis and latent variable modeling were conducted. Therefore, initially a large set of items was reduced via item analysis and confirmatory factor analysis (CFA) [49].

3.1 Ethics

The study was conducted in accordance with the requirements of the local ethics committee at our university. These requirements include, among other things, avoiding unnecessary stress, excluding risk and harm, and anonymizing participants. The personal data collected was limited to age, gender, and education. Particularly sensitive data (e.g. ethnicity, religion, health data) was not collected. Participants were not mislead but transparently informed about the procedure and goals of the study. They also had the option to end the survey at any time without giving a reason. They subsequently gave their informed consent to participate. *Respondi*, as the panel provider, is ISO-certified [35], complies with the GDPR, and ensures that the personal data is stored separately from survey data. Finally, the data analysis was conducted completely anonymized.

3.2 Item Phrasing and Questionnaire Design

Items were developed based on the established security questionnaire HAIS-Q [55] as well as recommendations of the German Federal Office for Information Security (BSI) on how to secure one's computer, smartphone, and data [12]. The goal here was to draw on established topics, while at the same time covering a wide variety of practically relevant topics. Building on the theoretical considerations and the envisaged dimensions of the final questionnaire, items were developed separately with regard to privacy knowledge and privacy behavior. To ensure accuracy, plausibility and content validity – the representative depiction of the characteristics to be investigated – we had the quality of the items reviewed by three experts in the field of privacy research.

Privacy knowledge items assessed knowledge about a broad spectrum of privacy issues, such as "The only data that are stored by me on the Internet are those that I have given myself". The theoretical considerations for these items were to ask for specific factual knowledge on best practices for privacy protection and potential threats, without surveying the participants' actual behavior. Items were answered on a 5 point Likert scale, ranging from *1 – I disagree* to *5 – I strongly agree* [61]. In general, higher agreement indicated more pronounced

privacy knowledge. However, items with reversed polarity were also included in the survey in order to prevent typical response biases that frequently occur in surveys.

Privacy behavior items assessed behavioral patterns in various contexts, such as "I have adjusted the privacy settings in social media so that I disclose less personal data". The theoretical considerations for the behavior items consisted of asking about specific privacy-relevant behaviors that may be performed in participants' daily lives. Again, items were answered on a Likert scale, ranging from *1 – never* to *5 – always*. Additionally, for all items the option to answer "I do not understand the question" was provided. For privacy behavior, too, higher answering scores indicated a more pronounced privacy behavior in general.

In total, 39 items were created for privacy knowledge and 42 items for privacy behavior (see Table 17). The items were developed in line with the principles of psychometrically sound item generation so that they would be, among other things, concise, non-suggestive, uni-dimensional, and understandable [49]. Before finalizing the questionnaire, N = 12 laypersons answered the questionnaire online within a pretest and confirmed that all developed items were comprehensible.

3.3 Sample

A representative online survey with German citizens was conducted in January 2020, using *LimeSurvey* and the ISO-certified [35] panel provider *Respondi*. *Respondi* provided the sample (N = 1,091) which was matched to the distribution of age, gender, income, region, and education according to the general German population during the data collection by using corresponding quotas (see Tab. 15 for sample information).

The sample was reduced by controlling for quality check questions such as requests to mark a specific answering box so that the sample size for the the initial item reduction based on item characteristics was N = 820. For the CFA, the sample was split half to enable a cross-validation. For the iterative item reduction (see Sec. 3.4), incomplete data points had to be disregarded due to the estimation method used in the CFA. This resulted in a larger number of data points the lesser the number of items that remained in the model. By this we aimed to make optimal use of information within the training sample, without dropping whole data points for single missing item responses. Consequently we had an effective sample size of N = 332 for the final model in the training set, and N = 324 for the cross-validation. These data points, summing up to N = 656, were used to predict factor scores and to assign a privacy persona.

3.4 Item Reduction and Evaluation of the Factor Structure of the Questionnaire

In order to reduce the overall item set, initially all items were analyzed concerning their range, item difficulty, and discriminatory power. The range should include all five points of the answering scale. In order to only keep items with a certain ability to differentiate between participants, the interquartile range was applied as a variance criterion for ordinally scaled data. In this course, items with an interquartile range of less than two were excluded. For item difficulty a range from Pi = 20-80 was accepted, while for discriminatory power a range of the correlation of r = .3 - .7. was accepted following the recommendations of Moosbrugger and Kelava [49].

To further reduce items and evaluate the theorized factor structure of the questionnaire, we performed a CFA using *R* and the package *lavaan*. To enable a cross-validation of the final CFA model at a later point, the sample was randomly split in half at this point, a training set and a test set. As the data consisted of ordinal Likert-scale items, the asymptotically distribution free method Weighted Least Squares with Adjustments for Means and Variances (WLSMV) estimation was applied. This approach leads to a mean and variance adjusted chi-square statistic and correspondingly scaled, robust fit indices. Importantly, this approach is regarded to be robust with ordinal responses, since the factor-analytical model does not decompose the covariance matrix of the observed variables but the matrix of their polychoric correlations [33]. The models were constrained so that the items were only allowed to load on the theorized factors they were each conceptualized for.

Likelihood ratio test or further likelihood based criteria (Akaike information criterion (AIC), Bayesian information criterion (BIC)) were not available for the present non-nested WLSMV approach in the CFA. Therefore, we focused on measures of local fit (modification indices (MIs), standardized residuals (SRs)) to explore weaknesses of the model and further reduce the item set. This approach was based on the literature, which proposed that items which are associated with several large MIs and SRs are rather nonspecific and dropping them generally eliminates multiple strains in the model [10]. Thus, we removed items which were associated with two or more outlying MIs (> 4 [37]) and SRs (> 2.58 [10]). To assess local fit in the final models

more closely, a threshold of 1.96 for SRs was considered as an indicator of remaining areas of poor fit [10]. This was done in an iterative process: First, an initial CFA model was constructed based on the initially reduced item set based on range, item difficulty, and discriminatory power (see Figure 1 for an overview). Then, items associated with two or more outyling MIs and SRs were iteratively removed. If several items were associated with two or more outlying MIs and SRs, the one with the largest sum of MIs was removed. After removal of this item, another CFA model was calculated and the local fit was assessed again. Because at least three items are recommended within a scale [58], we prevented an item from being removed, if its removal would have left only two items within a scale. In this case, not the item with multiple outlying MIs and SRs itself was removed, but the other item associated with this poor local fit. This iterative process stopped when no more items showed two or more clearly outlying MIs and SRs. It is important to note, that the use of MIs in this context is not uncontroversiial. They are based on chi-square statistics which can be used to informally compare models, but not to conduct formal chi-square difference tests [42]. Our goal in this process was not to artificially increase model fit but to be rather strict and rely on at least some objective criteria (together with SRs), since common likelihood-based criteria were not available in our case. To control for the risk of potential overfitting in this process, a cross-validation was performed on an independent sub-sample.

The global model fit was evaluated by using chi-square test, Root Mean Square Error of Approximation (RMSEA), Comparative Fit Index (CFI), Tucker-Lewis Index (TLI), and Standardized Root Mean Square Residual (SRMR), respectively their robust versions. Common recommendations considered for evaluating the fit indices were close to or below .08 for RMSEA, close to or above .95 for CFI and TLI, as well as close to or below .08 for SRMR [34, 42]. A significant chi-square test generally indicates that a proposed model fails to reproduce the polychoric correlation matrix. To test the hypothesis of close model fit we used robust RMSEA and its 90%-confidence interval as a criterion for evaluating the closeness-of-fit. Here, we considered the model results a close fit if the lower boundary of the confidence interval was <.05 [64]. In addition to the model fit, reliability was assessed using ordinal α [23]. Additionally, we provided ω as a measure of congeneric reliability [59]. Convergent and discriminant validity were assessed using average variance extracted (AVE) and heterotrait-monotrait ratio of correlations (HTMT) [31].

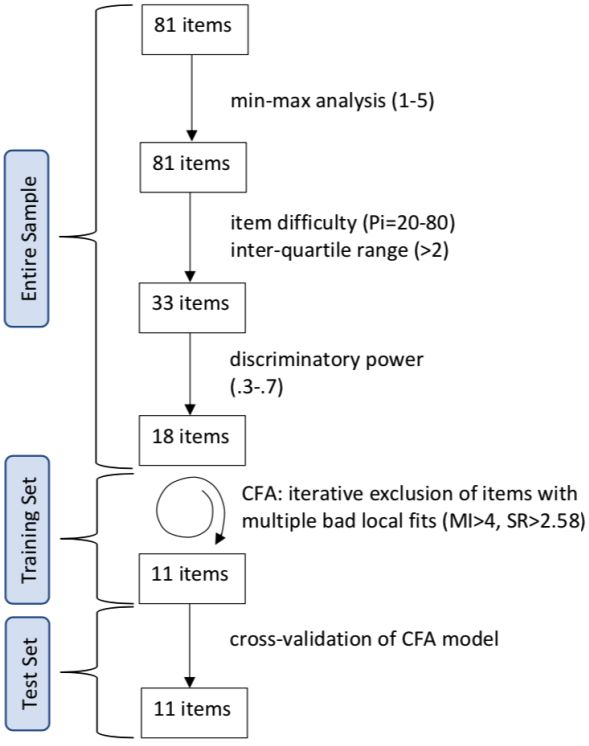

Fig. 1. Methodological approach.

In order to test the stability of the resulting CFA model, we divided the sample into a training set (N = 410) and a test set (N = 410) prior to conducting the CFA. The initial model was established within the training set based on the previously described criteria. For the cross-validation, the resulting parameter estimates for factor loadings of the initial model were used to fit another CFA model on the test set.

3.5 Clustering and Statistical Analysis

Based on the factor analysis, factor scores were calculated for the two latent variables privacy knowledge and privacy behavior by applying the fitted model with the `lavPredict()` function on the observation matrix, This function uses the thresholds (τ) for the ordinal answer options to estimate the latent response value for each item and predicts the score values by computing the weighted sum of these values for each factor with the factor loadings (λ) as weights. The goal then was to segment the user space into different personas. Here, a data-driven clustering method was applied on the entire sample. The advantage of this approach is that there is no need to specify a predefined number of clusters, but rather the optimal number of clusters results

from the combination of the data-driven proposed clustering boundaries and theoretical considerations [26]. Specifically, hierarchical agglomerative clustering with the Ward-method was used [73]. Agglomerative clustering describes the approach of initially considering each data point as its own cluster and iteratively combining data points into a common cluster should they be sufficiently similar based on the dissimilarity/distance measure used. We applied the distance measure of *squared euclidean distance* between data points, which is common for Ward's method [26]. The number of clusters was determined based on the visual analysis of the dendrogram in combination with theoretical considerations. The dendrogram presents a graphical overview of the hierarchical relationships between data points, i.e. how (dis)similar groups of data points are depending on the number of clusters that are chosen (see Appendix 6).

To evaluate the cluster solutions, we performed ordinal logistic regression analyses with the factor scores for privacy knowledge and behavior as predictors. Based on these, the goal was to predict membership to a specific persona cluster. To gain some independence from the initial covariance matrix, this regression was performed within the subsample not used for building the initial CFA model (test set). Importantly, the clustering is not considered to provide any evidence on the robustness of the CFA model. Instead, it is an application for assessing which personas can be uncovered in the data.

Finally, demographic differences between the privacy persona clusters were analyzed within the entire sample. To do this, ordinal logistic regression models were calculated using gender, age, and education as predictors for membership to one of the privacy persona clusters. Subsequent exploratory analyses within significant demographic factors we conducted using the Kruskal-Wallis test (due to issues with non-normality of residuals and unequal variances between groups) with Dunn's post-hoc test (corrected for multiple comparisons with a Benjamini-Hochberg adjustment [7]).

4 Results

This section presents an initial item analysis (section 4.1), followed by a CFA (section 4.2.1) and cross-validation (section 4.2.3) to examine the factor structure of the questionnaire. Subsequently, a cluster analysis is presented to identify distinct privacy personas (section 4.3) and the personas are characterized (section 4.4).

4.1 Item Analysis

The aim here was to exclude obviously unsuitable items that did not meet the basic criteria in terms of a psychometrically valid questionnaire. A min-max analysis revealed that for all items, the full range of answering possibilities (1-5) was used (see Appendix 5). To evaluate the average agreement to items, the item difficulty was analyzed and only items with a difficulty index between Pi = 20-80 remained. A visual analysis of the distributions of responses revealed that some items led to highly skewed responses, with most of the participants giving similar answers. Therefore, items with an interquartile range of less than two were excluded. Finally, the discriminatory power of items was calculated in order to evaluate how well a specific item differentiates between individuals. This correlation of each item with the other items of the respective theorized scale was calculated using the `psych` package for *R* and items with a discriminatory power outside of the range between .3 to .7 were excluded. Performing these steps of the item analysis resulted in an item pool of five items for the privacy knowledge scale and thirteen items for the privacy behavior scale. Thus, the initial item set was reduced to a total of 18 items for the subsequent evaluation of the factor structure of the questionnaire.

4.2 Evaluating the Factor Structure of the Questionnaire

4.2.1 Confirmatory Factor Analysis (H1)

To validate the two factors derived from the theoretical considerations, the presumed model was evaluated with a CFA with the two correlated factors privacy behavior and privacy knowledge. This model was built within a training set (N=410) to allow for a cross-validation of the model on a validation set (N = 410). Since only cases without missing data for the remaining items could be kept, the effective samples were slightly smaller (training set: N = 332, validation set: N = 324). The original model was constructed according to the theoretical rationale so that privacy knowledge items loaded on the first factor and privacy behavior items loaded on the second factor. The CFA model was estimated using the respective function of the `lavaan` package in *R* for ordinal data with polychoric correlations. WLSMV was used as estimation method, incomplete cases were omitted, and latent variable variances were fixed to 1.0.

The analysis of local fit resulting from the initial CFA model revealed some areas of poor fit. Since the goal was to develop a concise questionnaire, our approach was rather strict and we therefore dropped items which were rather non-specific. In this course, we iteratively removed items associated with several outlying MIs and SRs (see section 3.4 for details) and reran the CFA to reassess the local fit. This way we dropped seven clearly unspecific items and arrived at eleven items in the final set. There potentially exists a risk that reliance on purely empirical criteria has a chance component to it. However, in this process no complex new pathways were added but single items were removed. Thus, the removal was always in line with theoretical considerations (i.e. items only loaded on the factors they were constructed for). The corresponding modification indices and associated expected parameter chance values of the removed items can be found in Table 12. To control for the risk of overfitting, a cross-validation was performed on an independent sub-sample (section 4.2.3).

With this approach, the hypothesis (H1) of a close fit of the final CFA model with a two-factor structure could be confirmed, based on the 90% confidence interval of the (robust) RMSEA (0.000-0.046). The chi-square test was non-significant, suggesting that the model successfully reproduced the polychoric correlation matrix (p=.836). The other (robust) fit indices also suggested a satisfactory global fit in accordance with commonly considered values for CFI, TLI and SRMR (see Table 2). It is important to note, that the thresholds for the indicators differed vastly between items (see Table 7). This is not surprising considering the skewed distribution of some items (see Fig. 5) but can bias the measurement model. For example, it was shown that larger differences in thresholds might lead to biased ω estimates, especially in smaller samples [75]. While the threshold differences might have had an negative effect on the model properties, it also highlights the importance of an robust estimator method, such as WLSMV that we used in the model, as opposed to non-robust estimators as maximum likelihood. Nevertheless this could point to the fact that the items were not ideal.

In terms of local fit, the inspection of MIs and SRs still indicated occasional points of problematic fit in the solution (highest MI: 3.63, highest SR: 2.39) (see Table 9). In three instances, there remained critical SRs between indicators of the factors privacy knowledge and behavior which were mainly associated with the item K2. In these cases, the CFA model both underestimated and overestimated associations between K2 and several other indicators (B1, B5, B6). Even though there were no clearly outlying values visible, these indicated misloadings raised initial doubts about the accuracy of the questionnaire's presumed two-factor structure. The consequences thereof are further illuminated in the cross-validation (section 4.2.3). Furthermore, the two factors correlated strongly ($r = .68$, $p < .001$), which is in line with the theoretical considerations. A complete description of model specifications, including the thresholds for the polychoric correlations and SRs can be found in the Appendix (7 & 9). The final behavior (B) and knowledge (K) items are listed in Table 6. The final model and the corresponding factor structure of the questionnaire are depicted in Fig. 2. The eight behavior items as well as the three knowledge items show varying factor loadings from .33 to .86 with regard to the corresponding scales. On this basis, factor scores were calculated for each subject serving as a measure for the relative privacy knowledge and behavior. Taken together, based on the RMSEA confidence interval, the hypothesis of a close fit could be confirmed for the two-factor model of the questionnaire (H1).

4.2.2 Reliability and Validity

To be generally considered a reliable scale, α and $\omega > .70$ would be required [28] while convergent validity would be established with an average variance extracted (AVE) of $> .50$ [22]. For the present questionnaire, ordinal alpha as a measure for reliability was $\alpha = .78$ for the subscale privacy behavior and $\alpha = .55$ for the subscale privacy knowledge. Congeneric reliability was $\omega = .73$ for the subscale privacy behavior and $\omega = .53$ for the subscale privacy knowledge. To assess convergent validity the AVE was calculated, yielding .32 for the privacy behavior scale and .33 for the privacy knowledge scale. Thus, both validity and reliability could not be established for the self-developed scale. Discriminant validity, however, could be established using the heterotrait-monotrait correlation ratio (HTMT) and yielded a value of .63 - which was in the required range below .9 [31].

The weaknesses in reliability and validity went hand in hand with below-average factor loadings. Only two items (B8 & K1) showed higher factor loadings than .7. Four items (B2, B4, B5 & B7) showed factor loadings just above the generally recommended lower limit of .4. K3 showed a factor loading of .33, thus explaining only about 10 % of the variance. Consequently, a substantial amount of error variance remained when looking at the single items. This lead to an overall reduced reliability of both scales. Furthermore, since K3 explained only

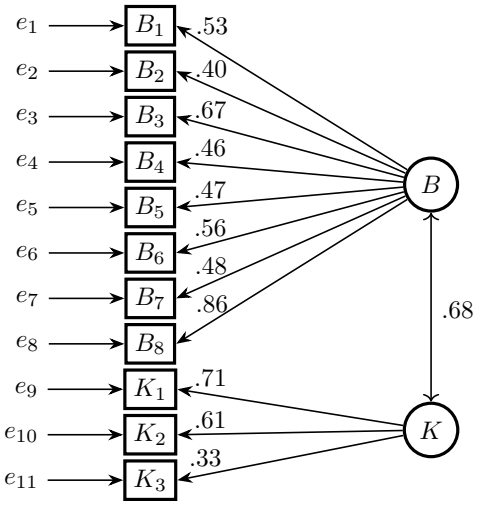

Fig. 2. Factor structure and loadings of the CFA model which were also used as constraints for the cross-validation model.

about 10 % of the variance in the case of the knowledge subscale, the knowledge factor was primarily estimated using only the two items K1 & K2. However, a stable model ideally consists of at least three indicators [58]. Thus, the lower loadings of a number of items reduced the stability of the model which negatively affected the overall reliability and validity of the model. Taken together, the low reliability and low AVE indicate that both factors knowledge and behavior extracted fairly little signal and a substantial amount of unexplained variance remained in the two-factor model. As a consequence, the low signal-to-noise ratio can attenuate the relationship with other variables [27].

4.2.3 Cross-Validation (H2)

The CFA model was cross-validated on the test set. Here, the parameter estimates for the factor loadings of the initial model were used to fit the CFA model on the new data set and obtain the residuals and fit measures. The hypothesis (H2) of a close fit in an independent sub-sample could be confirmed for this model, too. The lower boundary of the 90% confidence interval of the robust RMSEA was below .05 (0.044-0.074). The chi-square test was significant suggesting that the model did not successfully reproduce the polychoric correlation matrix ($p < .001$). The other (robust) fit indices also suggested a worse fit than the original CFA model. While they were either only slightly above a recommended threshold (SRMR) or close to the range of an acceptable fit (CFI, TLI), all these thresholds were exceeded and taken together they contradict a similarly good fit for the cross-validation model as for the training model (see Table 2). Again, the thresholds for the indicators differed vastly (see Table 10). Furthermore, the inspection of MIs and SRs revealed further points of poor fit (highest MI: 18.14 with several more above 4, highest SR: 3.74 with several more above 2.58). For example, B4 showed particular weaknesses and was associated with two high SRs (3.74 & 3.64) indicating a poor local fit. Additionally, B3 showed several significant SRs. These include positive residuals with two items of the knowledge subscale. This indicated that the item was rather unspecific and might also have loaded on the other factor. Combined with some negative SRs of B3, but also with several more significant SRs suggesting underestimation and overestimation of associations between indicators, this cast doubt on the adequacy of the presumed two-factor structure. This might reflect that the theoretical approach of assigning the items to their respective scale based on face validity could have been flawed. The unidimensionality of the subscales could therefore not be unequivocally assumed.

In terms of reliability, ordinal alpha was $\alpha = .73$ for the subscale privacy behavior and $\alpha = .55$ for the subscale privacy knowledge. Congeneric reliability was $\omega = .73$ for the subscale privacy behavior and $\omega = .54$ for the subscale privacy knowledge. The AVE was .32 for the privacy behavior scale and .33 for the privacy knowledge scale. Thus, the AVE remained below the generally accepted threshold of .5 so that convergent validity could not be established in the cross-validated model as well. Discriminant validity, on the other hand, could be established using the heterotrait-monotrait correlation ratio (HTMT) which yielded a value of .69. It is important to note that the cross-validation does not provide substantive additional evidence for the validity and reliability of the scale. This is because the factor loadings were constrained in the cross-validation model and the corresponding measures are based on the loadings and the error variances. Our primary objective here was to ensure that no overfitting occurred in the training model and to provide a validation of the fit measures.

Table 2. Robust measurements of model fit for the training (1) and cross-validated (2) CFA model (based on mean and variance adjusted chi-square).

Model	X^2 / df (p-value)	RMSEA	CFI	TLI	SRMR
(1)	51.12 / 43 (.185)	.024	.992	.990	.040
(2)	114.90 / 54 (<.001)	.059	.931	.930	.081

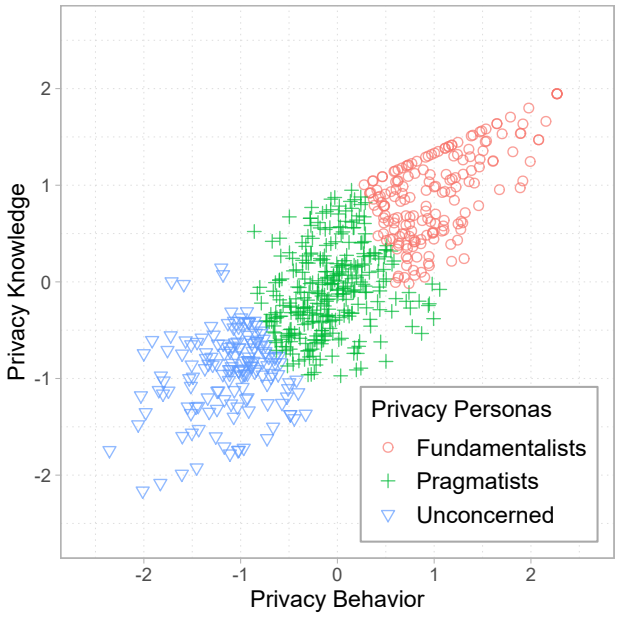

Fig. 3. Overview of Privacy Personas based on factor scores for privacy knowledge and behavior.

4.3 Cluster Analysis (H3)

Based on the factor scores resulting from the CFA, the goal was to segment users on the privacy knowledge-behavior ratio by applying a combination of data-driven methods and theoretical considerations. The results of the clustering were evaluated based on the resulting dendrogram (see Appendix 6). Here, the relative dissimilarity between possible clusters (depending on the number of clusters chosen) was used as an indicator for the most plausible number of clusters. Based on the dendrogram, both a three-cluster and a four-cluster solution seemed reasonable since the relative increase in dissimilarity between three or four clusters was within a reasonable range. This observation was combined with the findings that the scales of privacy knowledge and behavior were strongly correlated. Since subjects did not show a high variability on both scales and some weaknesses existed with regard to the reliability of the questionnaire, a too fine-grained distinction seemed arbitrary. For this reason, we adopted the three-cluster solution and proceeded with mapping corresponding privacy personas. The three clusters can be seen in Fig. 3 and are distributed quite uniformly along the dimensions of privacy knowledge and behavior.

To confirm that privacy knowledge and behavior could actually predict cluster affiliation, we performed an ordinal logistic regression analysis using the `orm()` function from the `rms` package for *R*. The regression model estimates showed significant main effects for both predictors (Table 3). Thus, privacy knowledge and behavior scores could indeed be used to predict which cluster a subject would fall into. Taken together, the hypothesis that different privacy personas would be revealed by a cluster analysis could be confirmed (H3).

4.4 Characteristics of Privacy Personas (H4)

The three-cluster solution turned out to be distributed quite uniformly along the dimensions of privacy knowledge and behavior. Our goal was to use persona designations that were as neutral as possible. Since the initial privacy segmentation by Westin used relatively neutral specifications and the number of Westin's personas matches the results of our clustering, we concluded that there is no need to invent new personas. As our results thematically matched Westin's designations, the following personas were defined in our study:

(1) Fundamentalists scoring high in privacy knowledge and behavior,
(2) Pragmatists scoring average in privacy knowledge and behavior,
(3) Unconcerned scoring low in privacy knowledge and behavior.

While these class names were inspired by Westin's work, no concurrent validation with Westin's personas was conducted. Hence, there is no empirical evidence for equivalency or correlation of these personas. Since previous research indicated that general differences in privacy knowledge and behavior exist depending on demographic factors, we analyzed whether such differences run along the line of our privacy personas. To test the hypothesis that age, gender, and education have an effect on membership of privacy persona, we conducted an ordinal regression analysis with these factor as predictors and the persona membership as the dependent variable. The model estimates can be found in Table

Table 3. Ordinal logistic regression modeling the effects of privacy behavior and knowledge scores on the assigned privacy persona. $\chi^2(2) = 837.7$, $p = <.001$

Predictor	$\hat{\beta}$	SE	$Wald\ Z$	p
Behavior score	22.37	3.17	7.07	<.001
Knowledge score	13.95	2.04	6.85	<.001

Table 4. Regression model for demographic variables. The categorical variables are dummy coded, with gender (male) and education (other) as reference. $\chi^2(9) = 101.57$, $p = <.001$

Predictor	$\hat{\beta}$	SE	$Wald\ Z$	p
Gender (female)	0.16	0.16	1.02	.31
Age	0.05	0.01	8.64	<.001
Education:				
German "Hauptschulabschluss"	2.23	1.38	1.63	.10
German "Mittlere Reife"	2.36	1.39	1.70	.09
Completed vocational training	1.81	1.36	1.32	.19
Univ. of appl. sc. entr. qualific.	2.49	1.40	1.79	.07
Higher education entr. qualific.	1.97	1.38	1.42	.15
Higher education	1.67	1.37	1.21	.22

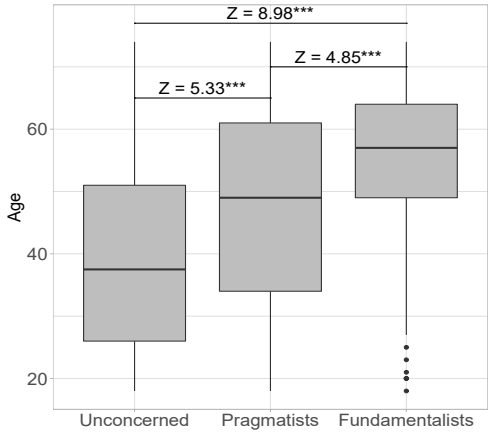

Fig. 4. Age across privacy persona groups. Z is the test statistic of Dunn's post hoc test. *** indicates a significant group difference at the level $p < .001$

4 and show that only age could significantly predict the personas. Subsequent exploratory analyses revealed that privacy personas significantly differed in age, $H(2) = 80.71$, $p < .001$. Pairwise comparisons showed significant age differences between all personas (see Figure 4). Among them, *Fundamentalists* were the oldest (median = 57 years), followed by *Pragmatists* (median = 49 years) and *Unconcerned* (median = 37.5 years). Table 5 shows the distributions of the personas over their demographics indicating that gender and education are fairly evenly distributed among the three personas.

5 Discussion

The goal of this study was to identify typical privacy personas and for this purpose, to conduct a methodical evaluation of the self-developed concise questionnaire with the subscales privacy knowledge and privacy behavior. Using item analysis and confirmatory factor analysis to evaluate the construct validity, the initial item set of 81 items was reduced to eleven items with a satisfactory global model fit (H1). This was generally supported by subsequent cross-validation on a separate sub-sample, albeit with weaker evidence (H2). Moreover, the models revealed severe weaknesses with regard to local fit, validity and reliability. The resulting factors privacy knowledge and privacy behavior turned out to be correlated and were used for a subsequent data-driven clustering of subjects. This clustering led to the distinction between three user groups: (1) *Fundamentalists* scoring high in privacy knowledge and behavior, (2) *Pragmatists* scoring average in privacy knowledge and behavior, and (3) *Unconcerned* scoring low in privacy knowledge and behavior (H3). Subsequent analysis of potential differences in demographic variables revealed age to be a significant predictor for persona membership, while gender and education were not (H4).

5.1 Questionnaire Evaluation

5.1.1 Methodological Aspects

The initially established two-factor model reflecting the questionnaire's two scales showed a satisfactory global fit. The corresponding closeness-of-fit hypothesis based on RMSEA could be confirmed, too. A subsequent cross-validation of this established model showed a worse fit based on the global fit indices (CFI, TLI, SRMR). Furthermore, problems with poor local fit emerged that had already been foreshadowed in the training model, but were most apparent in the cross-validation. The implied ambiguity represented via significant SRs cast doubt on the unidimensionality of the two subscales, on the accuracy of the presumed two-factor structure and thus, on the construct validity. These results are in line with literature emphasizing the limited meaningfulness of global fit indices alone [42]. Taken together, the acceptable global fit measures alone are not sufficient to justify the adequacy of the assumed factor structure.

Further weaknesses emerged during the evaluation of the questionnaire. These are mutually dependent and mainly concern the magnitude of the factor loadings, the convergent validity and the reliability. These issues are partly associated with an imperfect local fit of the models. Only the global fit indices showed satisfactory values, while the local fit evaluation revealed remaining areas of suboptimal fit, especially in the cross-validation

Table 5. Median ages and contingency table of privacy personas for gender and education.

	Unconcerned	Pragmatists	Fundamentalists
Age [years] (median)	37.50	49.00	57.00
Gender			
Male	82	171	97
Female	78	140	88
Education			
Other	1	1	0
German "Hauptschulabschluss"	32	86	54
German "Realschulabschluss"	6	25	15
Completed vocational training	39	88	61
Univ. of appl. sc. entrance qualification	12	18	14
Higher education entrance qualification	31	33	18
Higher education	38	59	23

sample. The factor loadings ranged from .33 to .71. This issue of partially low factor loadings was mirrored in the AVE as a measure for convergent validity. Ideally, the items of one factor should be strongly interrelated and explained well by the underlying factor [10]. The AVE for the two factors privacy knowledge and privacy behavior, however, was 0.33 and 0.32. Thus, on average, more variance remained in the error of the items than in the variance explained by the latent construct (privacy knowledge or behavior). The two scales thereby missed the standard criterion, that the items should account for at least of the factor variance. Similarly, the reliability did not show consistently satisfactory values for both scales. While ω for the factor privacy behavior was satisfactory (.73), ω for the factor privacy knowledge was clearly below (.53) the necessary threshold to be considered reliable (>.70). The differences in reliability are plausible in view of the fact that a higher number of items is generally associated with higher reliability [60].

The low AVE and low reliability indicate that the respective items do not consistently and comprehensively measure the same, narrowly circumscribed construct. Conversely, the latent constructs privacy knowledge and privacy behavior do not influence the item responses to the desired extent. This issue might be caused in part by the fact that some items may have actually been too ambiguous and broad, resulting in imperfect descriptions of the constructs privacy knowledge and behavior. Thus, it cannot be ruled out that some items also unintentionally targeted overlapping constructs.

While the global fit indices indicate sufficient model fit, the observed local fit problems and low signal-to-noise ratio at the item level ultimately cast doubt on the stability of the two-factor model. As described earlier, the assessment of model fit should not be based solely on a favorable global fit. On the one hand, the unidemensionality of the subscales is challenged by local fit issues that indicate some items to also load on the other factor. On the other hand, the factor knowledge is essentially calculated only based on two items, since the third item (K3) only roughly explains 10 % variance. These issues emerge at the more global level of AVE and reliability and compromise the questionnaire's informational content.

These results also point to potential issues regarding the content validity of the items. While the behavior items were constructed to reflect specific privacy behaviors and the knowledge items to reflect knowledge on good privacy practice and potential vulnerabilities, the issues in local fit indicate that there might be cross-loadings between the theorized factor items. The problem could either lie in the theorized factors not reflecting the real world variables, or in the fact that on the item level, it is hard isolate the specific factor influence, since knowledge and behavior are so closely related that one might argue one is the premise for the other.

5.1.2 Lessons Learned and Recommendations for Future Research

The results of this study provide important insights for future research. In general, the approach taken demonstrates the importance of precise item development in relation to a narrowly circumscribed construct presumed behind it. Furthermore, only a thorough methodological evaluation of aspects such as factor structure, reliability and validity enables an accurate assessment of the quality criteria of a questionnaire. The lack of such a thorough methodological approach, however, can potentially lead to contradictory empirical results when using a questionnaire - particularly, when the core prob-

lem is actually the imperfect questionnaire itself. This is particularly important in light of the fact that even established and widely used instruments such as the IUIPC reveal weaknesses when thoroughly examined for their psychometric properties [27]. Despite psychometrical weaknesses of the questionnaire in the present study, we hope that this work contributes to foster the consideration of important methodological aspects for dealing with self-developed questionnaire items.

These include (1) the consideration of ordinality of Likert item responses. The WLSMV estimator used in the present study adequately accounts for this and should be considered. It should be noted, however, that in the context of item reduction and choice between competing models, common likelihood-based information criteria are not available with WLSMV but only with ML estimators. We therefore relied on empirical respecification based on MIs and SRs which presents its own problems. MIs are primarily used to modify hierarchical models [42]. However, the removal of items as in the present study is a non-nested change. Besides SRs, these MIs were used as one criterion for removing items. While chi-square values can be used to informally compare models, the chi-square difference test cannot be applied [42]. While we did not directly apply a chi-square difference test between reduced and non-reduced models, MIs themselves are based on chi-square statistics. Further common criteria such as AIC and BIC on the other hand are also likelihood based and primarily used for ML estimators but not for WLSMV. Therefore, the item reduction method did not rely on widely accepted criteria. And indeed, the problems related to local fit as well as low validity and reliability showed that the item reduction process and final model were not beyond doubt. However, the cross-validation confirmed to some extent that items were not removed solely on the basis of chance, otherwise we would have expected overfitting to occur and consequently a significantly worse model fit for the validation data. The ideal model comparison approach for non-nested models estimated with WLSMV thus remains an open question and should be considered when choosing the estimator.

Furthermore, (2) a sound cross-validation of models should be carried out, especially if changes are made to the original model. Otherwise, there is a risk of artificially inflating model fit, which might not be transferable to other samples. As discussed above, our model selection approach bears the risk of overfitting, contributing to the importance of this step. Also, instead of fitting a new CFA model with the same factor structure to our validation sample, we reused the loadings of our training model to prevent the inflation of fit measures.

Moreover, (3) the construct validity of questionnaires – also of rather established ones – should be investigated and tested, e.g., with factor analyses. Importantly, local fit should be investigated in this context, too, as this may reveal circumscribed weaknesses in terms of misloadings and potential challenges to unidimensionality assumptions even if the global model fit looks good. We conceptualized our questionnaire on a presumed two factor model based on the literature and face validity of the items. Our results may suggest that this assumption was incorrect, and therefore, such assumptions should be questioned and reviewed critically. In this course, it is advisable to precede the CFA with an exploratory factor analysis (EFA), even if items were developed specifically for two scales as in the present study. This ensures that surprising or rather unexpected problems, weaknesses, and ambiguities in individual item formulations may become visible at an early stage. This should be considered when determining the sample sizes, to ensure independent samples for EFA, CFA and cross-validation.

Furthermore, (4) the extent to which generalizable conclusions about empirical associations and descriptions of privacy personas are drawn must be directly related to the transparently reported reliability and validity coefficients, and thus the quality, of the questionnaires used. They represent quantifications of the signal-to-noise ratio whereby a low signal-to-noise ratio attenuates relations with other factors and can result in shaky, contradicting and generally not well founded results.

Finally, (5) transparent reporting of all relevant information (e.g. reliability and validity coefficients, standardized residuals, thresholds of the CFA, covariance matrices) helps other researchers to assess the quality of used questionnaires and supports further validation efforts, which should therefore be highly encouraged.

Overall, due to the weaknesses in terms of convergent validity and reliability well below the generally accepted thresholds, the present questionnaire should not be used in its current form. These result in limited sensitivity to accurately discriminate between personas due to the low signal-to-noise ratio. Nevertheless, future research could benefit from our results by factoring in the challenges we encountered during its development process. Assuming a sound statistical foundation, such a concise questionnaire could provide relevant insights, especially if there is no capacity for a more thorough assessment at that time. Therefore, the present results should be considered as a starting point for a thorough

evaluation of the items and instruments used in studies - especially when no established instruments are used.

5.2 Characteristics of Privacy Personas

The present study confirms the close relationship of privacy knowledge and behavior as evidenced by the strong correlation of the two factors. Consequently, a diversion between privacy knowledge and behavior was not observed in any of the persona groups. Such an effect, where knowledge does not necessarily predict corresponding behavior, however, has been observed in other studies [5, 70]. One experimental study, for example, demonstrated that technically-skilled and privacy-aware users do not necessarily avoid potential privacy intrusions to a greater extent than less-skilled users [5]. However, this particular study only considered the download and use of a mobile app and the privacy implications in a narrower sense rather than looking at a broad overview of privacy behaviors. In another experimental field study, a mobile application was developed with the goal to educate users in order to increase privacy awareness and knowledge [24]. This approach led participants to improve privacy conditions on their smartphones and more actively inform themselves about privacy related topics in general. Thus, the strong relationship between privacy knowledge and behavior has also been confirmed in a study with a high ecological validity.

In this context, however, the question arises, to what extent privacy knowledge and behavior are actually distinct in terms of their content validity and how well they can be assessed using self-reports. Obviously, there is an association between the two: a certain level of knowledge about privacy issues is the basis for being able to engage in certain privacy behaviors. However, the analysis of discriminant validity still suggested that the two scales of the questionnaire indeed measure two different constructs. This is in line with the theoretical assumption of our study, according to which separate items were developed targeting the knowledge domain on the one hand and the behavioral domain on the other. With regard to privacy behavior, it has to be considered that the questionnaire can only assess self-reports and not actual privacy behavior. This is where the privacy paradox, the divergence between attitudes and actual behavior [1, 44] becomes relevant as this may also bias self-reports of privacy behavior. The effects of the privacy paradox can only really be avoided if actual behavior is analyzed. We attempted to minimize the negative effects of self-reports by formulating items that target predominantly specific behaviors rather than mere behavioral intentions and attitudes.

In terms of demographic characteristics of privacy personas, only age significantly predicted persona membership in the present study. The findings that older individuals are more privacy-sensitive is thereby in line with other studies. For example, studies have highlighted that older adults (>40) are more concerned about privacy than younger adults [72] and less likely to self-disclose information on social media [40]. As for gender and education, the present study did not find any evidence that these are predictors of privacy personas, although some studies have considered them relevant moderators in the privacy domain [9, 51, 54, 71]. However, the poor reliability and validity of the instrument must be taken into account when interpreting the demographic differences. Against this background, these results contain a considerable amount of error and should not be regarded as universally valid.

5.3 Relation to Previous Privacy Segmentations

The results of this study join a number of studies that attempt to conceptualize privacy personas [19, 21, 50, 65]. However, unlike other approaches, our focus also lied on the methodological evaluation of the self-developed questionnaire. This evaluation revealed significant weaknesses in terms of validity and reliability, which is why the description of privacy personas can only be considered a very rough approximation. The early segmentation by Westin with three privacy personas has been widely used in other studies [45]. When criticism arose due to the weak scientific foundation of this segmentation [57], other attempts resulted in a different number of privacy personas (see Table 5). More differentiated, i.e. five personas were found using the dimensions of privacy knowledge and motivation [19], whereas less differentiated, i.e. three personas were found using the dimensions of privacy attitudes and behavior [65]. One influential factor for the potential to discover highly differentiated privacy personas is the correlation between the considered scales. The correlation between the two scales in the present study reduced the potential for finding rather “contradictory” personas who, for example, score high in privacy knowledge but show rather weak privacy behavior. Interestingly, in another study privacy attitudes and behavior were also correlated and the segmentation also resulted in only three privacy personas [65]. However, the survey

methodology is generally limited here, since individuals with high privacy knowledge might tend to report accordingly high privacy behavior and are not willing to reveal contradictory behavior. In contrast, a segmentation leading to five privacy personas reported no significant correlation between the dimensions of privacy knowledge and motivation. Therefore, the question of how many privacy personas there "really" are may be impossible to answer. Instead, the number of privacy personas might be best conceptualized as a function of the privacy dimensions under consideration and the degree to which they are related.

Our approach also puts the criticism of Westin's threefold segmentation in a different light. The main criticism here was that the items used to assign the personas did not stand up to rigorous scientific scrutiny [41, 57]. Subsequent trials to confirm this threefold segmentation using correlations with behavioral scenarios often failed to provide evidence for the practical usefulness of the segmentation [15, 41, 74]. This, however, might not be the result of the non-existence of three distinct privacy personas but rather of the false attribution to the corresponding personas due to the use of a suboptimal instrument. The present results highlight the challenges associated with developing a robust instrument for clustering privacy personas - which is a prerequisite for accurate descriptions of privacy personas. In fact, the heterogeneity in the number of privacy personas described in the literature could be partly explained by the differing quality of the instruments used. Importantly, it was never our goal to specifically look for Westin's personas, but rather to see which segmentation appears most plausible based on our concise questionnaire. It is important to note here, that only the class names are inspired by Westin's work, while no actual correlation to his classification was evaluated in the present study. The threefold segmentation appeared most plausible in the context of our data but other segmentations would have been theoretically possible, for example, if a fivefold segmentation would have described the data better.

5.4 Limitations

Some limitations of this study need to be taken into consideration. First, the results are based on the participants' self-reported privacy behavior which is not necessarily fully congruent with their actual behavior. The discrepancy between intentions and actual behavior has been reported before [1] and represents a general limitation of the survey methodology. In terms of the clustering, it should be noted that the number of personas was determined by a visual analysis of the corresponding dendrogram. Since no quantitative threshold was used in this process, the "true" number of clusters therefore remains uncertain to some degree. In addition, weighted factor scores were calculated for privacy knowledge and behaviors, rather than simply summing or averaging item responses. Such factor scores are dependent on the specific parameter estimates resulting from the CFA and therefore may not be fully generalizable to other samples. However, this adequately accounted for the greater variability in factor loadings between different items.

6 Conclusion

Privacy persona segmentations promise to structure the digital user space to better understand individual user needs and provide individualized support. One practical hurdle in providing individualized support is the lack of concise and valid instruments with which privacy personas can be found. Therefore, the present study used a self-developed questionnaire with eleven items on the two dimensions of privacy knowledge and privacy behavior in order to assign individuals to one of the three privacy personas (1) *Fundamentalists*, (2) *Pragmatists*, and (3) *Unconcerned*. The evaluation of characteristics of the three privacy personas suggested that simply looking at social group membership (e.g., by gender or education) is sometimes too superficial. Instead, it seems a more sensitive approach to segment users according to their actual privacy knowledge and privacy behavior rather than demographic characteristics. However, the thorough evaluation of the questionnaire revealed clear weaknesses in terms of validity and reliability coefficients, which are well below required thresholds in order for it to be considered a sound instrument. Thus, the present results can only be seen as a starting point for further advancing sound instruments to determine privacy personas, considering the lessons learned and presented in this study. Overall, the approach of this study sheds light on the fact that it is generally advisable to thoroughly evaluate questionnaires in terms of item characteristics and their underlying structure. A failure to do so might be one reason for the heterogeneity in the number of privacy personas found in the literature. Ultimately, the description of privacy personas can only be as accurate as the instruments used to determine them are methodologically sound.

7 Acknowledgements

This research work has been funded by the German Federal Ministry of Education and Research and the Hessian Ministry of Higher Education, Research, Science and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE and by the Deutsche Forschungsgemeinschaft (DFG) – SFB 1119 (CROSSING) – 236615297 as well as RTG 2050 (Privacy and Trust for Mobile Users) – 251805230.

References

[1] A. Acquisti and J. Grossklags. Privacy Attitudes and Privacy Behavior. *Economics of Information Security. Advances in Information Security*, 12:165–178, 2004. 10.1007/1-4020-8090-5_13.

[2] T. Alashoor and R. Baskerville. The privacy paradox: The role of cognitive absorption in the social networking activity. In *2015 International Conference on Information Systems: Exploring the Information Frontier, ICIS 2015*, dec 2015. URL https://aisel.aisnet.org/icis2015/proceedings/SecurityIS/5.

[3] G. Bansal. Distinguishing between privacy and security concerns: An empirical examination and scale validation. *Journal of Computer Information Systems*, 57(4):330–343, 2017.

[4] 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, oct 2019. 10.2478/popets-2019-0066.

[5] S. Barth, M. D. de Jong, M. Junger, P. H. Hartel, and J. C. Roppelt. Putting the privacy paradox to the test: Online privacy and security behaviors among users with technical knowledge, privacy awareness, and financial resources. *Telematics and Informatics*, 41(February 2019):55–69, 2019. ISSN 07365853. 10.1016/j.tele.2019.03.003.

[6] L. Baruh and Z. Cemalcilar. It is more than personal: Development and validation of a multidimensional privacy orientation scale. *Personality and Individual Differences*, 70:165–170, nov 2014. ISSN 01918869. 10.1016/j.paid.2014.06.042.

[7] Y. Benjamini and Y. Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995. ISSN 2517-6161. 10.1111/j.2517-6161.1995.tb02031.x. URL https://www.jstor.org/stable/2346101.

[8] A. Bergström. Online privacy concerns: A broad approach to understanding the concerns of different groups for different uses. *Computers in Human Behavior*, 53:419–426, jul 2015. ISSN 07475632. 10.1016/j.chb.2015.07.025.

[9] T. Biselli and C. Reuter. On the Relationship between IT Privacy and Security Behavior: A Survey among German Private Users. In *16th International Conference on Wirtschaftsinformatik*, pages 1–17. Association for Information Systems, feb 2021. URL https://aisel.aisnet.org/wi2021/NInformation12/Track12/3.

[10] T. A. Brown. *Confirmatory factor analysis for applied research*. The Guilford Press, New York, NY, 2014.

[11] 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, jan 2007. ISSN 15322890. 10.1002/asi.20459.

[12] Bundesamt für Sicherheit in der Informationstechnik (BSI). Das Internet sicher nutzen: 10 Tipps zur sicheren Nutzung des Internets Sicher im digitalen Alltag. Technical report, BSI, 2021. URL https://www.bmi.bund.de/SharedDocs/topthemen/DE/topthema-cybersicherheit/cybersicherheit-broschure-internet.pdf?__blob=publicationFile&v=1.

[13] H. Cho, B. Knijnenburg, A. Kobsa, and L. I. Yao. Collective privacy management in social media: A cross-cultural validation. *ACM Transactions on Computer-Human Interaction*, 25(3):1–33, jun 2018. ISSN 15577325. 10.1145/3193120.

[14] H. Cho, P. Li, and Z. H. Goh. Privacy Risks, Emotions, and Social Media: A Coping Model of Online Privacy. *ACM Transactions on Computer-Human Interaction*, 27(6), nov 2020. ISSN 15577325. 10.1145/3412367. URL https://doi.org/10.1145/3412367.

[15] S. Consolvo, I. E. Smith, T. Matthews, A. LaMarca, J. Tabert, and P. Powledge. Location disclosure to social relations: Why, when, & what people want to share. In *Conference on Human Factors in Computing Systems*, pages 81–90, 2005. ISBN 1581139985.

[16] R. E. Crossler and F. Bélanger. The Mobile Privacy-Security Knowledge Gap Model: Understanding Behaviors. In *Proceedings of the 50th Hawaii International Conference on System Sciences (2017)*, pages 4071–4080. University of Hawaii at Manoa, AIS IEEE Computer Society Press, 2017. ISBN 9780998133102. 10.24251/hicss.2017.491.

[17] T. Dinev and P. Hart. Internet privacy concerns and their antecedents – measurement validity and a regression model. *Behaviour and Information Technology*, 23(6):413–422, nov 2004. ISSN 0144929X. 10.1080/01449290410001715723.

[18] P. Dourish and K. Anderson. Collective information practice: Exploring privacy and security as social and cultural phenomena. *Human-Computer Interaction*, 21(3):319–342, sep 2006. ISSN 07370024. 10.1207/s15327051hci2103_2.

[19] J. L. Dupree, R. Devries, D. M. Berry, and E. Lank. Privacy personas: Clustering users via attitudes and behaviors toward security practices. In *Conference on Human Factors in Computing Systems - Proceedings*, pages 5228–5239, New York, NY, USA, may 2016. Association for Computing Machinery. ISBN 9781450333627. 10.1145/2858036.2858214.

[20] S. Egelman and E. Peer. The myth of the average user: Improving privacy and security systems through individualization. In *Proceedings of the 2015 New Security Paradigms Workshop*, volume 08-11-Sept, pages 16–28, New York, New York, USA, sep 2015. Association for Computing Machinery. ISBN 9781450337540. 10.1145/2841113.2841115.

[21] I. Elueze and A. Quan-Haase. Privacy Attitudes and Concerns in the Digital Lives of Older Adults: Westin's Privacy Attitude Typology Revisited. *American Behavioral Scientist*, 62(10):1372–1391, sep 2018. ISSN 15523381.

10.1177/0002764218787026.

[22] C. Fornell and D. F. Larcker. Evaluating Structural Equation Models with Unobservable Variables and Measurement Error. *Journal of Marketing Research*, 18(1):39–50, nov 1981. ISSN 0022-2437. 10.1177/002224378101800104. URL https://journals.sagepub.com/doi/abs/10.1177/002224378101800104.

[23] A. M. Gadermann, M. Guhn, and B. D. Zumbo. Estimating ordinal reliability for likert-type and ordinal item response data: A conceptual, empirical, and practical guide. *Practical Assessment, Research and Evaluation*, 17(3):1–13, nov 2012. ISSN 15317714. 10.7275/n560-j767. URL https://scholarworks.umass.edu/pare/vol17/iss1/3.

[24] N. Gerber, P. Gerber, H. Drews, E. Kirchner, N. Schlegel, T. Schmidt, and L. Scholz. FoxIT: Enhancing mobile users' privacy behavior by increasing knowledge and awareness. In *ACM International Conference Proceeding Series*, page 53–63. Association for Computing Machinery, 2018. ISBN 9781450363570. 10.1145/3167996.3167999.

[25] N. Gerber, B. Reinheimer, and M. Volkamer. Investigating People's Privacy Risk Perception. *Proceedings on Privacy Enhancing Technologies*, 2019(3):267–288, jul 2019. 10.2478/POPETS-2019-0047.

[26] P. Giordani, M. B. Ferraro, and F. Martella. *An Introduction to Clustering with R*. Springer, 2020. ISBN 9789811305528. 10.1007/978-981-13-0553-5_1.

[27] T. Groß. Validity and reliability of the scale internet users' information privacy concerns (iuipc). *Proceedings on Privacy Enhancing Technologies (PoPETs)*, 2021(2):235–258, 2021. doi:10.2478/popets-2021-0026. URL https://doi.org/10.2478/popets-2021-0026.

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

[29] 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. 10.2478/popets-2020-0020.

[30] K. Hartwig and C. Reuter. Nudging Users Towards Better Security Decisions in Password Creation Using Whitebox-based Multidimensional Visualizations. *Behaviour & Information Technology (BIT)*, 2021. 10.1080/0144929X.2021.1876167.

[31] 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, aug 2015. ISSN 15527824. 10.1007/s11747-014-0403-8.

[32] F. Herbert, G. M. Schmidbauer-Wolf, and C. Reuter. Differences in IT Security Behavior and Knowledge of Private Users in Germany. *Proceedings of the International Conference on Wirtschaftsinformatik (WI)*, pages 168–184, 2020. 10.30844/wi_2020_v3-herbert.

[33] F. P. Holgado–Tello, S. Chacón–Moscoso, I. Barbero–García, and E. Vila–Abad. Polychoric versus Pearson correlations in exploratory and confirmatory factor analysis of ordinal variables. *Quality and Quantity*, 44(1):153–166, jan 2010. ISSN 15737845. 10.1007/s11135-008-9190-y.

[34] D. Hooper, J. P. Coughlan, and M. R. Mullen. Structural equation modelling: guidelines for determining model fit. In *Journal of Business Research Methods*, volume 6, pages 53–60, 2008.

[35] International Organization for Standardization (ISO). ISO 20252:2019: Market, opinion and social research, including insights and data analytics — Vocabulary and service requirements. Technical Report 01, ISO, 2019. URL https://www.iso.org/standard/73671.html.

[36] J. Isaak and M. J. Hanna. User Data Privacy: Facebook, Cambridge Analytica, and Privacy Protection. *Computer*, 51(8):56–59, aug 2018. ISSN 15580814. 10.1109/MC.2018.3191268.

[37] J. Jaccard and C. Wan. *LISREL Approaches to Interaction Effects in Multiple Regression*. Sage, Thousand Oaks, California, 1996. 10.4135/9781412984782.

[38] T. N. Jagatic, N. A. Johnson, M. Jakobsson, and F. Menczer. Social phishing. *Communications of the ACM*, 50(10):94–100, 2007. ISSN 00010782. 10.1145/1290958.1290968.

[39] C. Jensen, C. Potts, and C. Jensen. Privacy practices of Internet users: Self-reports versus observed behavior. *International Journal of Human Computer Studies*, 63(1-2):203–227, jul 2005. ISSN 10715819. 10.1016/j.ijhcs.2005.04.019.

[40] M. Kezer, B. Sevi, Z. Cemalcilar, and L. Baruh. Age differences in privacy attitudes, literacy and privacy management on Facebook. *Cyberpsychology*, 10(1), may 2016. ISSN 18027962. 10.5817/CP2016-1-2. URL https://cyberpsychology.eu/article/view/6182/5912.

[41] J. King. Taken out of context: An empirical analysis of Westin's privacy scale. In *Symposium on Usable Privacy and Security*, pages 1–8, 2014.

[42] R. B. Kline. *Principles and practice of structural equation modeling*. Fourth edition. New York : Guilford Press,, 2016. URL https://search.library.wisc.edu/catalog/9910110667902121.

[43] B. P. Knijnenburg. Privacy? I Can't Even! Making a Case for User-Tailored Privacy. *IEEE Security and Privacy*, 15(4):62–67, 2017. ISSN 15584046. 10.1109/MSP.2017.3151331.

[44] S. Kokolakis. Privacy attitudes and privacy behaviour: A review of current research on the privacy paradox phenomenon. *Computers and Security*, 64(November):122–134, 2017. ISSN 01674048. 10.1016/j.cose.2015.07.002.

[45] P. Kumaraguru and L. F. Cranor. Privacy Indexes: A Survey of Westin's Studies. Technical report, Institute for Software Research International, 2005.

[46] P. K. Masur, D. Teutsch, and S. Trepte. Entwicklung und Validierung der Online-Privatheitskompetenzskala (OPLIS). *Diagnostica*, 63(4):256–268, oct 2017. ISSN 2190622X. 10.1026/0012-1924/a000179.

[47] T. McGill and N. Thompson. Gender differences in information security perceptions and behaviour. In *ACIS 2018 - 29th Australasian Conference on Information Systems*, pages 1–11, 2018. 10.5130/acis2018.co.

[48] R. E. Miller and J. Melton. A Typology of Student Social Media Users: A Posting Behavior Perspective. In *Proceedings of the Eleventh Midwest Association for Information Systems Conference*, 2016. URL http://aisel.aisnet.org/mwais2016/21.

[49] H. Moosbrugger and A. Kelava. *Testtheorie und Fragebogenkonstruktion*. Springer-Verlag, 2020. ISBN 978-3-662-61531-7. 10.1007/978-3-662-61532-4. URL http://

medcontent.metapress.com/index/A65RM03P4874243N.pdf.

[50] A. Morton and M. A. Sasse. Desperately seeking assurances: Segmenting users by their information-seeking preferences. In *2014 12th Annual Conference on Privacy, Security and Trust, PST 2014*, pages 102–111. Institute of Electrical and Electronics Engineers Inc., 2014. ISBN 9781479935031. 10.1109/PST.2014.6890929.

[51] D. O'Neil. Analysis of internet users' level of online privacy concerns. *Social Science Computer Review*, 19(1):17–31, feb 2001. ISSN 08944393. 10.1177/089443930101900103.

[52] G. Öğütçü, Ö. M. Testik, and O. Chouseinoglou. Analysis of personal information security behavior and awareness. *Computers and Security*, 56:83–93, feb 2016. ISSN 01674048. 10.1016/j.cose.2015.10.002.

[53] S. Pape, A. Ivan, D. Harborth, T. Nakamura, S. Kiyomoto, H. Takasaki, and K. Rannenberg. Re-evaluating internet users' information privacy concerns: the case in japan. *AIS Transactions on Replication Research*, 6(1):18, 2020.

[54] Y. J. Park. Do men and women differ in privacy? Gendered privacy and (in)equality in the Internet. *Computers in Human Behavior*, 50:252–258, sep 2015. ISSN 07475632. 10.1016/j.chb.2015.04.011.

[55] K. Parsons, D. Calic, M. Pattinson, M. Butavicius, A. McCormac, and T. Zwaans. The Human Aspects of Information Security Questionnaire (HAIS-Q): Two further validation studies. *Computers and Security*, 66:40–51, may 2017. ISSN 01674048. 10.1016/j.cose.2017.01.004. URL https://linkinghub.elsevier.com/retrieve/pii/S0167404817300081.

[56] S. Preibusch. Guide to measuring privacy concern: Review of survey and observational instruments. *International Journal of Human Computer Studies*, 71(12):1133–1143, 2013. ISSN 10959300. 10.1016/j.ijhcs.2013.09.002.

[57] S. Preibusch. Managing diversity in privacy preferences: How to construct a privacy typology. In *Symposium on Usable Privacy and Security*, pages 1–6, 2014. URL http://cups.cs.cmu.edu/soups/2014/workshops/privacy/s1p3.pdf.

[58] J. Raubenheimer. An item selection procedure to maximise scale reliability and validity. *SA Journal of Industrial Psychology*, 30(4), oct 2004. ISSN 0258-5200. 10.4102/sajip.v30i4.168.

[59] T. Raykov. Estimation of congeneric scale reliability using covariance structure analysis with nonlinear constraints. *British Journal of Mathematical and Statistical Psychology*, 54(2):315–323, nov 2001. ISSN 00071102. 10.1348/000711001159582.

[60] W. Revelle and D. M. Condon. Reliability from α to ω: A tutorial. *Psychological assessment*, 31(12):1395, 2019.

[61] B. Rohrmann. Empirische Studien zur Entwicklung von Antwortskalen für die sozialwissenschaftliche Forschung. *Zeitschrift für Sozialpsychologie*, 9:222–245, 1978.

[62] M. P. Rudolph and D. Feth. Usable Specification of Security and Privacy Demands: Matching User Types to Specification Paradigms. In *Mensch und Computer 2019 - Workshopband*. Gesellschaft für Informatik e.V., 2019. 10.18420/muc2019-ws-302-05.

[63] L. H. Rykkja, P. Lægreid, and A. L. Fimreite. Attitudes towards anti-terror measures: The role of trust, political orientation and civil liberties support. *Critical Studies on Terrorism*, 4(2):219–237, 2011. ISSN 17539153. 10.1080/17539153.2011.586206.

[64] K. Schermelleh-Engel, H. Moosbrugger, and H. Müller. Evaluating the Fit of Structural Equation Models: Tests of Significance and Descriptive Goodness-of-Fit Measures. *Methods of Psychological Research Online*, 8:23–74, 2003.

[65] E. M. Schomakers, C. Lidynia, and M. Ziefle. A Typology of Online Privacy Personalities: Exploring and Segmenting Users' Diverse Privacy Attitudes and Behaviors. *Journal of Grid Computing*, 17(4):727–747, dec 2019. ISSN 15729184. 10.1007/s10723-019-09500-3.

[66] K. B. Sheehan. Toward a typology of internet users and online privacy concerns. *Information Society*, 18(1):21–32, jan 2002. ISSN 01972243. 10.1080/01972240252818207.

[67] K. B. Sheehan and M. G. Hoy. Dimensions of privacy concern among online consumers. *Journal of Public Policy and Marketing*, 19(1):62–73, apr 2000. ISSN 15477207. 10.1509/jppm.19.1.62.16949.

[68] S. Sheng, M. Holbrook, P. Kumaraguru, L. F. Cranor, and J. Downs. Who falls for phish? A demographic analysis of phishing susceptibility and effectiveness of interventions. In *Conference on Human Factors in Computing Systems*, volume 1, pages 373–382, 2010. ISBN 9781605589299. 10.1145/1753326.1753383.

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

[70] J. M. Stanton, K. R. Stam, P. Mastrangelo, and J. Jolton. Analysis of end user security behaviors. *Computers and Security*, 24(2):124–133, 2005. ISSN 01674048. 10.1016/j.cose.2004.07.001.

[71] S. Tifferet. Gender differences in privacy tendencies on social network sites: A meta-analysis. *Computers in Human Behavior*, 93:1–12, apr 2019. ISSN 07475632. 10.1016/j.chb.2018.11.046.

[72] E. Van den Broeck, K. Poels, and M. Walrave. Older and Wiser? Facebook Use, Privacy Concern, and Privacy Protection in the Life Stages of Emerging, Young, and Middle Adulthood. *Social Media and Society*, 1(2), dec 2015. ISSN 20563051. 10.1177/2056305115616149.

[73] J. H. Ward. Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, 58(301):236–244, 1963. ISSN 1537274X. 10.1080/01621459.1963.10500845.

[74] A. Woodruff, V. Pihur, S. Consolvo, L. Brandimarte, and A. Acquisti. Would a privacy fundamentalist sell their DNA for $1000... if nothing bad happened as a result? The Westin categories, behavioral intentions, and consequences. *SOUPS '14: Proceedings of the Tenth Symposium On Usable Privacy and Security*, 1:1–18, 2014. URL https://www.usenix.org/conference/soups2014/proceedings/presentation/woodruff.

[75] Y. Yang and Y. Xia. Categorical Omega With Small Sample Sizes via Bayesian Estimation: An Alternative to Frequentist Estimators. *Educational and Psychological Measurement*, 79(1):19–39, jan 2019. ISSN 15523888. 10.1177/0013164417752008. URL https://journals.sagepub.com/doi/10.1177/0013164417752008.

[76] M. Zeng, S. Lin, and D. Armstrong. Are all internet users' information privacy concerns (iuipc) created equal? *AIS Transactions on Replication Research*, 6(1):3–17, 2020.

A Appendix

Table 6. Final behavior (B) and knowledge (K) items in the questionnaire.

B1	I use technology to help me control my personal information.
B2	I lock my technical device (laptop, smartphone, etc.) when I am not actively using it.
B3	I actively read the data protection and privacy regulations before I register with an online service (such as Facebook).
B4	I disclose as little information about myself as possible on the Internet (e.g., no information about profession, addresses, date of birth).
B5	I immediately uninstall all programs on my tech devices that I don't need.
B6	I make sure to use https connections.
B7	I delete my browsing history.
B8	I actively protect my data.
K1	In case of violations of data protection and/or privacy, I know where to report them (e.g. police).
K2	The only data that are stored by me on the Internet are those that I have given myself.
K3	To protect my data, I should only use public networks via a VPN (= virtual private network) connection.

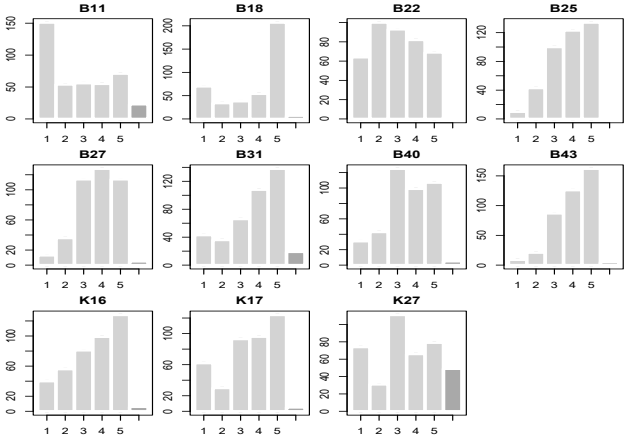

Fig. 5. Answer distributions of final item set (x-axis: Likert-Scale answers with the sixth option being "no answer", y-axis: nr. of answers).

Table 7. Parameter estimates for confirmatory factor analysis model. Parameters were estimated using WLSMV and ordinal item variables (s. thresholds), standard errors were estimated using "robust.sem".

Item	$\hat{\lambda}$	SE	z	p	Std.lv	Std.all
Behavior						
B11	0.528	0.050	10.621	0.000	0.528	0.528
B18	0.395	0.057	6.909	0.000	0.395	0.395
B22	0.674	0.039	17.099	0.000	0.674	0.674
B25	0.457	0.053	8.675	0.000	0.457	0.457
B27	0.473	0.051	9.324	0.000	0.473	0.473
B31	0.558	0.047	11.782	0.000	0.558	0.558
B40	0.483	0.050	9.558	0.000	0.483	0.483
B43	0.856	0.030	28.345	0.000	0.856	0.856
Knowledge						
K16	0.712	0.054	13.181	0.000	0.712	0.712
K17	0.611	0.054	11.216	0.000	0.611	0.611
K27	0.326	0.067	4.895	0.000	0.326	0.326
Covariances						
Latent factors	$\hat{\rho}$	SE	z	p	Std.lv	Std.all
B, K	0.678	0.053	12.846	0.000	0.678	0.678
Thresholds						
Item$\|\tau_i$	$\hat{\tau}$	SE	z	p	Std.lv	Std.all
B11$\|\tau_1$	-0.331	0.070	-4.704	0.000	-0.331	-0.331
B11$\|\tau_2$	0.023	0.069	0.329	0.742	0.023	0.023
B11$\|\tau_3$	0.436	0.071	6.117	0.000	0.436	0.436
B11$\|\tau_4$	0.879	0.079	11.056	0.000	0.879	0.879
B18$\|\tau_1$	-1.021	0.084	-12.204	0.000	-1.021	-1.021
B18$\|\tau_2$	-0.723	0.076	-9.529	0.000	-0.723	-0.723
B18$\|\tau_3$	-0.420	0.071	-5.900	0.000	-0.420	-0.420
B18$\|\tau_4$	-0.076	0.069	-1.096	0.273	-0.076	-0.076
B22$\|\tau_1$	-0.996	0.083	-12.019	0.000	-0.996	-0.996
B22$\|\tau_2$	-0.275	0.070	-3.941	0.000	-0.275	-0.275
B22$\|\tau_3$	0.331	0.070	4.704	0.000	0.331	0.331
B22$\|\tau_4$	0.971	0.082	11.831	0.000	0.971	0.971
B25$\|\tau_1$	-1.925	0.143	-13.488	0.000	-1.925	-1.925
B25$\|\tau_2$	-1.143	0.088	-12.986	0.000	-1.143	-1.143
B25$\|\tau_3$	-0.307	0.070	-4.377	0.000	-0.307	-0.307
B25$\|\tau_4$	0.478	0.072	6.657	0.000	0.478	0.478
B27$\|\tau_1$	-2.032	0.156	-13.028	0.000	-2.032	-2.032
B27$\|\tau_2$	-1.268	0.093	-13.587	0.000	-1.268	-1.268
B27$\|\tau_3$	-0.283	0.070	-4.050	0.000	-0.283	-0.283
B27$\|\tau_4$	0.574	0.073	7.839	0.000	0.574	0.574
B31$\|\tau_1$	-1.251	0.093	-13.518	0.000	-1.251	-1.251
B31$\|\tau_2$	-0.846	0.079	-10.757	0.000	-0.846	-0.846
B31$\|\tau_3$	-0.355	0.070	-5.031	0.000	-0.355	-0.355
B31$\|\tau_4$	0.387	0.071	5.466	0.000	0.387	0.387
B40$\|\tau_1$	-1.504	0.106	-14.160	0.000	-1.504	-1.504
B40$\|\tau_2$	-0.971	0.082	-11.831	0.000	-0.971	-0.971
B40$\|\tau_3$	-0.053	0.069	-0.767	0.443	-0.053	-0.053
B40$\|\tau_4$	0.619	0.074	8.371	0.000	0.619	0.619
B43$\|\tau_1$	-2.032	0.156	-13.028	0.000	-2.032	-2.032
B43$\|\tau_2$	-1.459	0.103	-14.106	0.000	-1.459	-1.459
B43$\|\tau_3$	-0.610	0.074	-8.265	0.000	-0.610	-0.610

Thresholds						
Item|τ_i	$\hat{\tau}$	SE	z	p	Std.lv	Std.all
B43|τ_4	0.205	0.069	2.957	0.003	0.205	0.205
K16|τ_1	-1.417	0.101	-14.032	0.000	-1.417	-1.417
K16|τ_2	-0.742	0.076	-9.736	0.000	-0.742	-0.742
K16|τ_3	-0.198	0.069	-2.848	0.004	-0.198	-0.198
K16|τ_4	0.436	0.071	6.117	0.000	0.436	0.436
K17|τ_1	-1.087	0.086	-12.649	0.000	-1.087	-1.087
K17|τ_2	-0.793	0.077	-10.251	0.000	-0.793	-0.793
K17|τ_3	-0.113	0.069	-1.644	0.100	-0.113	-0.113
K17|τ_4	0.512	0.072	7.088	0.000	0.512	0.512
K27|τ_1	-0.814	0.078	-10.454	0.000	-0.814	-0.814
K27|τ_2	-0.547	0.073	-7.518	0.000	-0.547	-0.547
K27|τ_3	0.236	0.070	3.395	0.001	0.236	0.236
K27|τ_4	0.783	0.077	10.148	0.000	0.783	0.783

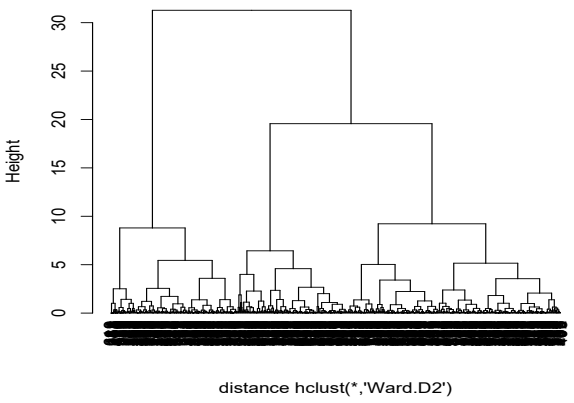

Fig. 6. Dendrogram as a result of hierarchical Ward clustering. At the beginning (the bottom), each individual represents a single cluster within the privacy knowledge-behavior ratio. Subsequently, more and more previously fine-grained clusters are combined into less fine-grained clusters on the way along the y-axis. The y-axis hereby represents the relative dissimilarity between the chosen cluster solution: the further along the y-axis, the less similar are the cases within the clusters.

Table 8. Model-implied (standardized) covariance matrix of the final CFA model in the training sample.

	B11	B18	B22	B25	B27	B31	B40	B43	K16	K17	K27
B11	1.00										
B18	0.21	1.00									
B22	0.36	0.27	1.00								
B25	0.24	0.18	0.31	1.00							
B27	0.25	0.19	0.32	0.22	1.00						
B31	0.29	0.22	0.38	0.26	0.26	1.00					
B40	0.25	0.19	0.33	0.22	0.23	0.27	1.00				
B43	0.45	0.34	0.58	0.39	0.40	0.48	0.41	1.00			
K16	0.25	0.19	0.32	0.22	0.23	0.27	0.23	0.41	1.00		
K17	0.22	0.16	0.28	0.19	0.20	0.23	0.20	0.35	0.43	1.00	
K27	0.12	0.09	0.15	0.10	0.10	0.12	0.11	0.19	0.23	0.20	1.00

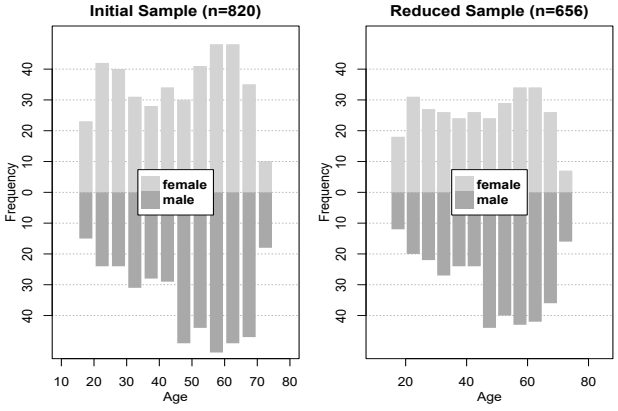

Fig. 7. Distribution of age and gender in the samples.

Table 9. Standardized model residuals from the fitted CFA model in the training sample (marked for significance at the alpha level 0.05(*). No residuals were significant at the alpha level 0.01).

	B11	B18	B22	B25	B27	B31	B40	B43	K16	K17	K27
B11	0.00										
B18	0.19	0.00									
B22	0.32	-0.12	0.00								
B25	-0.34	1.11	0.59	0.00							
B27	-0.23	0.91	*-2.11	0.21	0.00						
B31	0.73	-0.60	1.27	-1.55	-0.28	0.00					
B40	0.06	0.36	-1.27	1.56	*1.97	0.94	0.00				
B43	1.55	0.18	-1.29	0.31	-1.34	-0.55	*-2.00	0.00			
K16	-1.38	-0.18	1.54	-1.49	-0.45	0.77	-0.82	1.10	0.00		
K17	*-2.07	-1.11	0.82	-0.32	*2.09	*-2.39	-1.04	1.88	-0.31	0.00	
K27	-0.65	-1.81	0.48	-1.02	-0.47	0.81	0.24	0.76	-0.27	0.71	0.00

Table 10. Parameter estimates for cross-validation CFA. Item loadings λ_i were fixed and reused from the training model.

Covariances						
Latent factors	$\hat{\rho}$	SE	z	p	Std.lv	Std.all
B, K	0.656	0.064	10.281	0.000	0.656	0.656
Thresholds						
Item|τ_i	$\hat{\tau}$	SE	z	p	Std.lv	Std.all
B11|τ_1	-0.355	0.071	-4.982	0.000	-0.355	-0.355
B11|τ_2	0.062	0.070	0.887	0.375	0.062	0.062
B11|τ_3	0.599	0.074	8.039	0.000	0.599	0.599
B11|τ_4	1.018	0.085	12.034	0.000	1.018	1.018
B18|τ_1	-1.058	0.086	-12.308	0.000	-1.058	-1.058
B18|τ_2	-0.734	0.077	-9.531	0.000	-0.734	-0.734
B18|τ_3	-0.364	0.071	-5.092	0.000	-0.364	-0.364
B18|τ_4	-0.046	0.070	-0.666	0.506	-0.046	-0.046
B22|τ_1	-0.896	0.081	-11.070	0.000	-0.896	-0.896
B22|τ_2	-0.195	0.070	-2.772	0.006	-0.195	-0.195
B22|τ_3	0.517	0.073	7.064	0.000	0.517	0.517
B22|τ_4	0.980	0.083	11.752	0.000	0.980	0.980
B25|τ_1	-2.022	0.157	-12.917	0.000	-2.022	-2.022
B25|τ_2	-1.158	0.090	-12.909	0.000	-1.158	-1.158
B25|τ_3	-0.266	0.071	-3.768	0.000	-0.266	-0.266
B25|τ_4	0.414	0.072	5.751	0.000	0.414	0.414
B27|τ_1	-1.868	0.138	-13.524	0.000	-1.868	-1.868
B27|τ_2	-1.018	0.085	-12.034	0.000	-1.018	-1.018
B27|τ_3	-0.195	0.070	-2.772	0.006	-0.195	-0.195
B27|τ_4	0.599	0.074	8.039	0.000	0.599	0.599
B31|τ_1	-1.271	0.095	-13.434	0.000	-1.271	-1.271
B31|τ_2	-0.796	0.078	-10.156	0.000	-0.796	-0.796
B31|τ_3	-0.266	0.071	-3.768	0.000	-0.266	-0.266
B31|τ_4	0.323	0.071	4.541	0.000	0.323	0.323
B40|τ_1	-1.403	0.101	-13.834	0.000	-1.403	-1.403
B40|τ_2	-0.818	0.079	-10.362	0.000	-0.818	-0.818
B40|τ_3	-0.046	0.070	-0.666	0.506	-0.046	-0.046
B40|τ_4	0.491	0.073	6.737	0.000	0.491	0.491
B43|τ_1	-2.159	0.177	-12.215	0.000	-2.159	-2.159
B43|τ_2	-1.363	0.099	-13.737	0.000	-1.363	-1.363
B43|τ_3	-0.589	0.074	-7.931	0.000	-0.589	-0.589
B43|τ_4	0.266	0.071	3.768	0.000	0.266	0.266
K16|τ_1	-1.271	0.095	-13.434	0.000	-1.271	-1.271
K16|τ_2	-0.734	0.077	-9.531	0.000	-0.734	-0.734
K16|τ_3	-0.101	0.070	-1.442	0.149	-0.101	-0.101
K16|τ_4	0.448	0.072	6.190	0.000	0.448	0.448
K17|τ_1	-1.307	0.096	-13.564	0.000	-1.307	-1.307
K17|τ_2	-0.786	0.078	-10.053	0.000	-0.786	-0.786
K17|τ_3	-0.140	0.070	-1.996	0.046	-0.140	-0.140
K17|τ_4	0.491	0.073	6.737	0.000	0.491	0.491
K27|τ_1	-0.796	0.078	-10.156	0.000	-0.796	-0.796
K27|τ_2	-0.431	0.072	-5.971	0.000	-0.431	-0.431
K27|τ_3	0.282	0.071	3.989	0.000	0.282	0.282
K27|τ_4	1.071	0.086	12.398	0.000	1.071	1.071

Table 11. Model-implied (standardized) covariance matrix for the cross-validation CFA.

	B11	B18	B22	B25	B27	B31	B40	B43	K16	K17	K27
B11	1.00										
B18	0.21	1.00									
B22	0.36	0.27	1.00								
B25	0.24	0.18	0.31	1.00							
B27	0.25	0.19	0.32	0.22	1.00						
B31	0.29	0.22	0.38	0.26	0.26	1.00					
B40	0.25	0.19	0.33	0.22	0.23	0.27	1.00				
B43	0.45	0.34	0.58	0.39	0.40	0.48	0.41	1.00			
K16	0.25	0.18	0.31	0.21	0.22	0.26	0.23	0.40	1.00		
K17	0.21	0.16	0.27	0.18	0.19	0.22	0.19	0.34	0.43	1.00	
K27	0.11	0.08	0.14	0.10	0.10	0.12	0.10	0.18	0.23	0.20	1.00

Table 12. Modification indices (mi) and expected parameter change (epc) above the critical threshold of 4.00 considered for excluding items (in addition to the criterion of standardized residuals above the threshold of 2.58).

Excluded item (iteration i)	Relation	mi	epc
B29 (i=1)	B29 ∼∼ B30	14.43	0.23
	Knowledge =∼ B29	13.15	-0.48
	B29 ∼∼ K27	11.17	-0.22
	B29 ∼∼ B31	4.68	0.15
B21 (i=2)	B21 ∼∼ K16	19.33	0.29
	Knowledge =∼ B21	10.83	0.47
	B21 ∼∼ B28	8.39	-0.19
	B21 ∼∼ B31	4.71	-0.16
K38 (i=3)	Behavior =∼ K38	9.57	0.36
	B40 ∼∼ K38	7.45	0.16
	K27 ∼∼ K38	6.69	0.14
	K17 ∼∼ K38	5.62	-0.15
B30 (i=4)	B28 ∼∼ B30	9.93	-0.20
	B30 ∼∼ B31	6.74	0.16
	B11 ∼∼ B30	4.56	0.13
K40 (i=5)	K40 ∼∼ K16	11.01	-0.23
	B43 ∼∼ K40	5.58	0.13
	B31 ∼∼ K40	4.48	-0.13
B28 (i=6)	B25 ∼∼ B28	5.83	0.15
	B28 ∼∼ K27	5.13	-0.14
B13 (i=7)	B13 ∼∼ B18	6.60	0.16
	B13 ∼∼ B22	4.95	-0.14

Table 13. Standardized residuals for the cross-validation CFA (marked for significance at the alpha level 0.05(*) and 0.01(**)).

	B11	B18	B22	B25	B27	B31	B40	B43	K16	K17	K27
B11	0.00										
B18	0.28	0.00									
B22	-1.95	**-2.63	0.00								
B25	**-3.74	0.40	*-2.44	0.00							
B27	-0.53	0.46	*-2.40	-1.07	0.00						
B31	-0.33	0.00	-0.20	0.36	-1.30	0.00					
B40	-1.80	-1.43	*-2.24	0.44	*2.06	1.65	0.00				
B43	*-2.16	-1.41	-1.41	0.06	-1.24	-1.17	0.47	0.00			
K16	0.38	0.09	**2.79	-1.60	-1.99	0.22	0.56	0.42	0.00		
K17	*-2.51	*-2.52	0.70	-0.34	-0.09	-0.22	-0.39	*2.34	-0.46	0.00	
K27	-1.34	**-2.60	*1.96	**-3.64	-0.24	-1.88	-1.04	0.43	-1.49	2.33	0.00

Table 14. Variances & means (intercepts) for the final (standardized) CFA model in training sample (same for cross-validation)

Variances			
Item/factor	$\hat{\sigma}^2$	Means	Scale factor
B11	0.721	0.000	1.000
B18	0.844	0.000	1.000
B22	0.546	0.000	1.000
B25	0.791	0.000	1.000
B27	0.776	0.000	1.000
B31	0.688	0.000	1.000
B40	0.767	0.000	1.000
B43	0.267	0.000	1.000
K16	0.494	0.000	1.000
K17	0.627	0.000	1.000
K27	0.894	0.000	1.000
Behavior	1.000	0.000	1.000
Knowledge	1.000	0.000	1.000

Table 15. Demographical properties of the different subsamples used.

	Full Sample		Training		Validation	
Age group	n	%	n	%	n	%
18–24	81	10.0	34	10.3	27	8.4
25–34	131	16.1	48	14.5	58	18.1
35–44	115	14.1	53	16.1	45	14.0
45–54	159	19.5	72	21.8	60	18.7
55–64	196	24.1	67	20.3	84	26.2
65–75	132	16.2	56	17.0	47	14.6
Gender						
male	410	50.0	181	54.5	169	52.2
female	410	50.0	151	45.5	155	47.8
Education						
Other	3	0.4	0	0.0	2	0.6
German "Hauptschulabschluss"	221	27.0	84	25.4	88	27.2
German "Mittlere Reife"	58	7.1	21	6.3	25	7.7
Completed vocational training	231	28.2	95	28.7	93	28.8
Univ. of appl. sc. entr. qualific.	53	6.5	25	7.6	19	5.9
Higher education entr. qualific	103	12.6	40	12.1	42	13.0
Higher education	149	18.2	66	19.9	54	16.7
Total N	820		332		324	

Table 16. Sample refinement for different steps of analysis

Step	**N**
Initial sample	1,091
After exclusion of failed control items	820
Split-half sample (Training/Validation)	410
Initial model (NAs excluded)	280
item reduction iteration #1	286
item reduction iteration #2	290
item reduction iteration #3	304
item reduction iteration #4	326
item reduction iteration #5	326
item reduction iteration #6	326
item reduction iteration #7	332
Cross validation (NAs excluded)	324
Cluster analysis	656

Table 17. Complete item set used for creating the PriPel questionnaire (English translation).

B1	I protect my technical devices (smartphone, laptop, tablet, etc.) with a PIN/password
B2	I open emails only if I know the sender.
B3	I open email attachments even if I don't know the sender.
B4	I log in to my personal social media accounts via public Wi-Fi (during a train journey or in a café).
B5	I log in to my personal email accounts via a public Wi-Fi (during a train journey or in a café).
B6	I log in to my personal online banking accounts via a public Wi-Fi (during a train ride or in a café).
B7	I post numerous private photos on social media (Instagram, Facebook).
B8	I download any file I need on the internet without hesitation.
B9	I use a password manager.
B10	I use a privacy dashboard.
B11	I use technology to help me control my personal information.
B12	(Quality Control Item: Please choose answer option 2).
B13	I use messengers that have sufficient message encryption.
B14	I have adjusted the privacy settings in social media so that I disclose less personal data.
B15	I write down my passwords on a piece of paper.
B16	I only keep my passwords in my head.
B17	I have the password for my laptop on a piece of paper ready near my laptop.
B18	I lock my technical device (laptop, smartphone, etc.) when I am not actively using it.
B19	I use search engines like Google without hesitation (in terms of my privacy)
B20	I only provide as much information to the online service as is necessary (e.g. I do not fill in optional fields when creating an account).
B21	I report non-compliance with data protection and privacy regulations to the appropriate authorities (e.g. police, consumer protection agency, state data protection authorities).
B22	I actively read the data protection and privacy regulations before registering with an online service (e.g. Facebook).
B23	I use the same password for all my accounts.
B24	I update my technical devices immediately.
B25	I disclose as little information about myself as possible on the Internet (e.g. no details of profession, addresses, date of birth).
B26	When I sell or transfer my technical devices, I move all files and documents to the recycle bin and empty it afterwards.
B27	I immediately uninstall all programs on my tech devices thatI don't need.
B28	When surfing the Internet, I don't worry and open any page.
B29	I use different user accounts on my PC, only one of them has admin rights.
B30	When using public networks, I surf via a VPN (= virtual private network) connection.
B31	I make sure to use https connections.
B32	I have selected the WP2/WP3 encryption level for my router.
B33	I mainly use very short and easy-to-remember passwords.
B34	I share on Facebook when and where I go on holiday.
B35	I agree to cookies.
B36	(Quality Control Item: Please choose answer option 6).
B37	I open websites from links, for example, from emails.
B38	I post a lot on the Internet and delete it again if necessary.
B39	I use different email addresses for different online services (e.g. social media, email).
B40	I delete my browsing history.
B41	I am proficient in using the Internet.
B42	Data protection is important to me.
B43	I actively protect my data.
B44	When installing/starting up my technical devices (laptop, smartphone, etc.), I get help from an expert.

K1	I should protect my technical devices (smartphone, laptop, tablet, etc.) with a PIN/password.
K2	I can safely open emails from unknown senders.
K3	I can safely open attachments in e-mails from unknown senders.
K4	It is safe for me to log into my personal accounts (social media, banking, email) via public Wi-Fi (during a train journey or in a café).
K5	I think social media (Instagram, Facebook) is a good place to share my private photos.
K6	I can safely download any file on the internet if I need it.
K7	Technologies like Password Manager give me more control over my personal data.
K8	Communication via messengers like Whatsapp is sufficiently encrypted.
K9	You should adjust your privacy settings in social media so that less personal data is disclosed.
K10	Passwords must be sufficiently protected from third parties.
K11	I should always lock my technical device (laptop, smartphone, etc.) when I am not actively using it.
K12	(Quality Control Item: Please choose answer option 2).
K13	Using search engines like Google does not pose a threat to my privacy.
K14	I should be as sparing as possible in disclosing data on the Internet.
K15	In the event of violations of data protection or/and privacy, I must take action and report them.
K16	In case of violations of data protection and/or privacy, I know where to report them (e.g. police).
K17	The only data that is stored about me on the Internet is the data that I have given myself.
K18	There are e-mails with which scammers want to get you to reveal personal data.
K19	Data protection can be increased by not accepting cookies from third-party providers.
K20	I should always read data protection and privacy regulations before registering with an online service (such as Facebook).
K21	I should use the same password for all my accounts.
K22	I should regularly update the programs on my technical devices.
K23	Disclosing information on the internet can have consequences (e.g. disclosure of profession can allow assessment of credit-worthiness).
K24	Before I sell my technical devices to other people, I must delete all my data.
K25	I should uninstall unneeded programs from my technical devices.
K26	If I have installed an anti-virus program, I don't have to worry about surfing the Internet and can open any page I want.
K27	To protect my data, I should only use public networks via a VPN (= virtual private network) connection.
K28	I should only surf with a user account that has admin rights, that is safest.
K29	I should only use https connections in exceptional cases, they are not secure.
K30	(Quality Control Item: Please choose answer option 1).
K31	I don't need to encrypt my router.
K32	My passwords should be as short as possible so that I can remember them easily.
K33	When I go on holiday, I can share the holiday location and time on Facebook without hesitation. I may always agree to cookies on websites, they do not pose a (tracking) risk.
K34	I may always agree to cookies on websites, they do not pose a (tracking) risk.
K35	For security reasons, I should not open websites via links (e.g.: from e-mails).
K36	I can safely post anything on the Internet, as I can delete it again.
K37	I should use different e-mail addresses for different online services (e.g. social media, e-mail).
K38	I should always type in URLs (Internet addresses) manually so that my activities cannot be tracked.
K39	I should regularly delete my browsing history.
K40	I know what data of mine is used on the Internet. (never - always)
K41	I know how to shop safely on the Internet. (never - always)