"Half My Survey Data Is Bad; The Problem Is, I Can't Tell Which Half": Evaluating the Validatability of Large-Scale Survey Data Using 176 Algorithmically Generated Personas

Bernard J. Jansen
Department of Information Management
Peking University
Beijing, China
jjansen@acm.org

Marwan Akari
Qatar Foundation
Doha, Qatar
alakarimarwan@gmail.com

Safa Amin
Carnegie Mellon University in Qatar
Doha, Qatar
safa.amin2004@gmail.com

Soon-Gyo Jung
Qatar Computing Research Institute
Doha, Qatar
sjung@hbku.edu.qa

Danial Amin
School of Marketing and Communication
University of Vaasa
Vaasa, Finland
danial.amin@uwasa.fi

Joni Salminen*
School of Marketing and Communication
University of Vaasa
Vaasa, Finland
jonisalm@uwasa.fi

Abstract

Survey data is foundational to much user research, including design artifacts such as personas. However, if survey data is invalidatable, the credibility of any downstream analysis is fundamentally undermined. This work starts with the premise that distinguishing valid from invalid survey data solely through internal survey checks, such as attention checks, is practically infeasible. We conducted a large-scale empirical survey of social media users (N ≈ 20,000) and used persona creation as an analytical lens. Of these responses, 8,140 were classified as (ostensibly) valid and 11,860 as invalid based on passing or failing attention checks. We construct ten datasets by progressively replacing 'valid' responses with 'invalid' ones in increments of 10%. From each dataset, we generate 16 personas, resulting in a total of 176, and compare their divergence. Results were stable despite data degradation. Persona demographics remained unchanged in most conditions, with demographic consistency at nearly 90%; an average of 12% of persona attributes were unchanged. More than 80% of the survey item diversity was consistent across datasets. The findings are that large-scale survey data may be unverifiable through internal checks alone due to the Validity Masking Effect, and that artifacts in such data can obscure underlying data quality issues, highlighting the risks of relying solely on survey data.

CCS Concepts

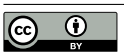
• **Human-centered computing**; • **Human computer interaction (HCI)**; • **Empirical studies in HCI**;

Keywords

survey data, data validity, automatic persona generation, data quality

ACM Reference Format:
Bernard J. Jansen, Marwan Akari, Safa Amin, Soon-Gyo Jung, Danial Amin, and Joni Salminen. 2026. "Half My Survey Data Is Bad; The Problem Is, I Can't Tell Which Half": Evaluating the Validatability of Large-Scale Survey Data Using 176 Algorithmically Generated Personas. In *37th ACM Conference on Hypertext (HT '26), September 14–18, 2026, London, United Kingdom.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3800935.3830835

*Corresponding author



HT '26, London, United Kingdom

ACM ISBN 979-8-4007-2564-7/2026/09
https://doi.org/10.1145/3800935.3830835

1 Introduction

Surveys are routinely applied in user research [48]. In human-computer interaction (HCI), surveys are prominent data sources for persona creation [42]. Personas are humanized abstractions of user data to inform design decisions [34] (see Figure 1). Personas are widely used in development and have been shown to outperform analytic approaches in specific design tasks [43]. However, this reliance on survey data masks a fundamental methodological difficulty: *the validity of the underlying survey data is rarely known, even when personas appear coherent, credible, and actionable.* Personas can further obscure inconsistencies, carelessness, or falsification in the survey data, creating an 'illusion of validity' [21]. Consequently, personas serve as a revealing exemplar and an analytic lens for a broader challenge that results appear empirically sound even when grounded in survey data whose validity is undetermined.

One aspect of persona creation that has received limited investigation is the validity of the survey data on which they are built. Therefore, generating personas [9] solely from survey responses raises concerns about accuracy, particularly as persona practice has shifted toward increasingly automated, data-driven, and algorithmic approaches that often 'distance' the persona creator from the underlying user data. Algorithmic persona generation (APG) uses machine learning (ML) techniques, such as clustering [7] and factorization [2], to create persona profiles from survey datasets.

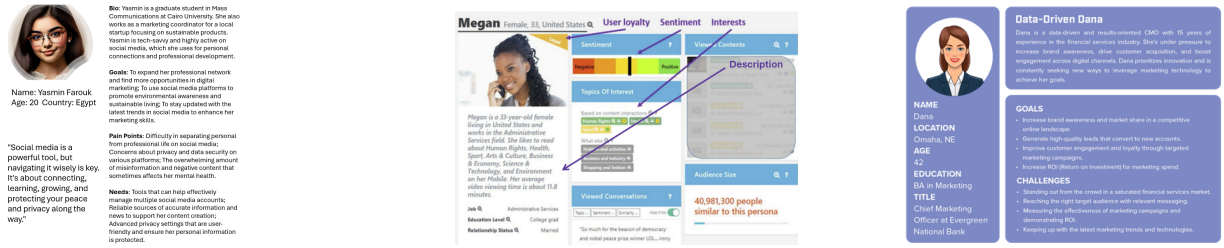


Figure 1: Examples of persona profiles, with an image, demographic information such as a name to humanize the data, and other attributes, such as pain points or needs (left), user loyalty, sentiment, and interests (center), and a cartoon persona (right).

Although these ML approaches offer scalability and rigor, when the underlying survey data is invalid, these APG techniques formalize data errors rather than correct them. Algorithmic sophistication cannot compensate for flawed data, and the resulting personas remain epistemically unstable despite any apparent coherence. This concern is further amplified by the increased reliance on online surveys, rather than interviews and focus groups, as the primary driver of the transition from manual to automated persona creation [48]. Survey data remains the dominant input for APG systems that leverage ML, artificial intelligence (AI), and data analytics [42].

Although data quality is recognized as essential, prior research has produced little empirical evidence on how invalid survey data affects analytical outcomes in the context of persona creation [40]. Standard survey practices often assume that internal mechanisms, such as attention checks, are sufficient for ensuring data quality; however, these controls do not resolve the deeper methodological problem that survey data itself must be verifiable [45]. When the data is unreliable [21], neither APG nor the personas it produces can be assumed to be trustworthy. As personas are frequently presumed to be derived from valid survey data [21], the consequences of invalid responses, their magnitude, persistence, and conditions of influence, remain largely unexplored. As a result, the "garbage in, garbage out" (GIGO) principle has largely been untested in persona and related research relying on surveys, leading to misplaced confidence in both survey validity and persona credibility [35]. Against this backdrop, we pose the following research questions (RQs) to investigate the difficulty of determining survey data validity using persona creation as an analytic lens:

RQ1: *Can changes in survey data quality be detected through shifts in the demographic characteristics of algorithmically generated personas?*

RQ2: *Can changes in survey data quality be detected through shifts in the persona profile attributes produced by algorithmic persona generation?*

RQ3: *Can changes in survey data quality be detected through shifts in the diversity of survey responses?*

First, RQ1 examines whether the integrity of survey data can be determined from the demographic stability of AGPs [27]. We formulate the following hypothesis: H1: *Increasing levels of invalid survey data are associated with increasing changes in persona demographics.*

Second, RQ2 investigates whether variations in survey data quality manifest as divergence in persona profile items, which encode core attributes of user segments and constitute a significant portion of the substantive content displayed in persona profiles [40]. We posit H2: *Increasing levels of invalid survey data are associated with increasing changes in persona profile attributes.*

Third, RQ3 investigates whether purported changes in survey data quality affect the diversity of survey responses, which serve as the basis for persona profile values. We note that this analysis focused on the survey data itself, so it is independent of any persona creation algorithm. We propose H3: *Increasing levels of invalid survey data are associated with increasing diversity in persona attributes.*

We address these RQs through an experimental design that reflects a realistic scenario in which invalid responses evade standard internal validity checks and are, therefore, treated as valid during analysis. As such, persona creation serves as an exemplar, demonstrating that current internal survey validity controls are insufficient for determining whether the data underlying survey-based artifacts can be trusted. To this end, these RQs and Hs address a fundamental issue: *whether survey data can be reliably trusted for research practices, such as persona creation* [37]. Although quantitative surveys often benefit from a 'mystique of numbers' that confers unwarranted credibility and objectivity [50], surveys are increasingly questionable given the emergence of large language models (LLMs) that can mimic human responses. If survey data is invalidatable, as this work assumes, then the personas and other artifacts derived from it are compromised. This risks poor design decisions and the creation of systems that fail to reflect actual user needs and behaviors. Accordingly, this study focuses on survey data reliability and its downstream consequences for persona creation [17], using personas as an exemplar of how invalid data can propagate through research processes undetected. The findings offer insights for researchers and practitioners on the limitations of existing survey validation methods, demonstrating that personas can appear resilient even when based on untrustworthy survey data.

2 Background

2.1 The Dilemma of Data Validity

Data quality is a foundational research construct [11], including personas derived predominantly from surveys. If survey data is unverifiable, the resulting personas are also unverifiable, undermining design and decision-making processes [15] as personas themselves

are difficult to verify. Prior work has demonstrated that inconsistencies in survey responses can hinder decision making efficiency [5]. It is documented that fraudulent inputs, including careless responding, falsified answers, and bot-generated data, are significant issues in survey research [28]. Although survey design strategies have been proposed [25], alongside algorithmic techniques aimed at detecting or mitigating low-quality data [22], none resolve the core methodological challenge: *invalid survey data often cannot be reliably distinguished from valid data using internal criteria alone.* These concerns are compounded by the growing presence of AI-generated survey responses [3] that can appear coherent and plausible, yet be substantively meaningless. This difficulty is exacerbated by LLMs, which can generate substantial volumes of responses that mimic valid human input while containing factual inaccuracies [24]. Accordingly, our research proceeds from the premise that, even with meticulous cleaning and preprocessing, internal survey-based validation is insufficient for establishing the validity of large-scale survey data or of the personas derived from it, positioning persona creation as a revealing exemplar of this broader limitation.

2.2 Personas and Data Quality

Although personas are frequently constructed from survey data, the problem of invalid survey responses has received surprisingly little direct attention in HCI persona research. Prior work addresses data quality concerns in principle but offers no practical guidance on how to determine whether data is low-quality. McGinn and Kotamraju [31] cautioned that poor data quality undermines persona accuracy and user experience (UX) yet offered no empirical analysis of how invalid data affects persona construction. Similarly, Matthews et al. [29] emphasized the risk of misleading personas, but focused on how practitioners perceive and use personas rather than on how invalid survey data propagates into persona artifacts. As a result, systematic empirical studies examining the impact of survey data on persona generation are rare, if not absent. The dominant implication of the "garbage in, garbage out" (GIGO) principle [12] is that low-quality input will visibly degrade outputs, discouraging closer scrutiny of cases in which artifacts remain plausible despite 'garbage' data. In contrast, our study investigates whether it is even possible to determine the validity of survey data for persona creation.

Advances in ML, AI, and data analytics (DA) are increasingly used to generate personas. Nevertheless, they do not resolve the fundamental problem of determining whether the underlying survey data is valid. Prior work has focused primarily on evaluating personas as artifacts rather than on validating the data from which they are derived. APG researchers typically focus on algorithms that enhance persona generation [56]. However, to our knowledge, these approaches have not examined how survey data quality itself influences persona outcomes or offered techniques for validating survey responses. Prior work in APG offers objectives for persona generation, including fairness, diversity, and consistency [41]. However, these objectives presuppose that the underlying survey data is valid, overlooking the possibility that personas may be grounded in unreliable or invalid responses. As a result, personas can appear reliable, yet they do not accurately represent the users they are intended to model. In our study, we examine how progressively higher levels of degraded survey data alter personas relative to supposedly valid data. As such, we demonstrate that personas built on invalidatable data appear trustworthy yet unverifiable, highlighting a critical challenge in HCI: user representation.

2.3 Previous Approaches to Data Validation for Personas

Related work proposes probabilistic approaches for assessing data source credibility [55], emphasizing trustworthiness through confidence, such as confidence modeling [4]. However, these approaches do not address the unresolved problem that, when survey data cannot be validated, neither personas nor other findings derived from it can be validated. This lack of empirical evidence on unverifiable survey data is the central focus of the present research, using persona generation as an example of the broader challenge of determining the validity of survey data in HCI. Chapman et al. [7] proposed a model for analyzing persona attributes, observing that persona descriptiveness diminishes as the number of attributes increases. However, even these critics of personas presume reliable input data, failing to address the effects of invalid survey responses. Similarly, Shenton et al. [16] advocate rigor in qualitative research, emphasizing credibility, transferability, dependability, and confirmability. However, these standards assume that the collected data is trustworthy and offer no mechanisms to verify validity. Salminen et al. [49] introduced evaluative dimensions, such as credibility, completeness, clarity, and consistency, for assessing personas. However, these metrics also operate at the level of the persona representation and remain agnostic to the quality of the survey data that generated them. These approaches reflect efforts to evaluate persona outputs, but they leave unresolved the central methodological challenge: when survey data quality is uncertain, personas still fail to accurately represent users.

Although prior work has acknowledged the importance of data quality, we are not aware of any study that directly examines the trustworthiness of survey data for persona creation or for other HCI applications aimed at understanding users. The validation of survey data quality is largely absent from prior HCI work, leaving a critical knowledge gap in understanding how data quality affects persona generation and other areas of research. Our study examines this challenge of distinguishing between valid and invalid survey data, highlighting a foundational risk in research and practices that rely on surveys.

3 Methodology

3.1 Materials

The dataset for this study was from a large-scale survey designed to capture user behaviors, preferences, and demographic characteristics relevant to the creation of personas for social media users in the Middle East and North Africa (MENA). The survey collected over 20,000 responses from 16 MENA countries through an international, reputable survey provider, in accordance with their company policy governing user data. Following the survey company's internal quality control procedures and two embedded attention-check items, 8,140 responses were classified as 'valid'. The remaining 11,860 responses were categorized as 'invalid' data, representing participants

who failed attention checks [32]. We use this valid-invalid separation as the foundation for systematically examining the validity of survey data and the implications for persona generation.

The survey incorporated two commonly used attention check methods to identify inattentive responding. The first was an instructional manipulation check embedded in the questionnaire that required respondents to select a specific response option to demonstrate that they had read the question carefully (i.e., "To show that you are actually reading the statements, select 'Disagree"'). The second attention check evaluated whether respondents followed simple instructions (i.e., "Select 'Strongly Disagree' to prove that you are actually reading the statements"). Both are standard methods employed in survey research. Online surveys frequently use these methods to detect invalid responses and represent standard internal survey quality controls [1]. To account for the possibility that earlier survey questions are less affected by attention, we split the demographic questions at the beginning (i.e., age) and end (i.e., gender, nationality) of the survey. Behavioral and preference items that were the primary basis for persona generation appeared throughout the survey. Demographic variables accounted for only a small subset of the 115 variables used in persona construction, with the majority derived from these behavioral and attitudinal items.

Our use of attention checks to separate the data does not imply that survey responses can be definitively classified as 'valid' or 'invalid' (our premise is that they do not). Instead, attention checks operationalize the dominant and widely accepted mechanism by which HCI and other survey researchers attempt to identify problematic responses (i.e., the invalid ones only). This framing of attention checks as an internal survey control check is intentional. If attention checks meaningfully distinguish between low-quality and high-quality survey data, then introducing responses that fail these attention checks should result in changes in downstream analyses, such as persona creation. However, if the commonly accepted internal validity survey control, relied upon by researchers, fails to produce detectable differences in survey data, this would raise concerns about its effectiveness as a validity check. The absence of change would challenge the practical effectiveness of these controls, raising concerns for artifact creation downstream.

To address this examination, this study analyzed data collected from a survey instrument comprising multiple items, supplemented with two attention-check questions. Several items were structured as matrix questions (e.g., "In the past week, approximately how much time per day have you spent on the following on average"), covering eleven major social media platforms such as Facebook, Twitter, YouTube, Snapchat, and Instagram. For analysis, each response within the matrix was treated as a distinct survey variable, aligning with the persona generation algorithm and standard survey research practices. This procedure produced 115 items used in the persona generation process. Before analysis, all responses, both 'valid' and 'invalid', based on passing or failing the attention checks, underwent the same pre-processing phase [8] involving standardization of categorical responses and Likert scale data. We make both datasets publicly available.[1]

[1]https://osf.io/en3gj/overview?view_only=415c7f8a5e2142c2b9c4751d6ef84e50

3.2 Experimental Conditions

We first constructed a baseline dataset of 'valid' survey responses, comprising 8,140 respondents who passed the attention checks. We designated it as the gold standard for the analysis. In survey-based studies, survey datasets that meet these criteria are assumed to be valid and suitable for downstream analytical tasks, including persona generation [57]. This gold-standard dataset serves as the reference condition for examining whether the accepted indicator of survey validity or invalidity yields discernible differences in persona outputs. We then construct ten additional datasets by introducing responses classified as supposedly 'invalid' into the baseline dataset in increments of 10%; each 10% increment replaced an equivalent proportion of reportedly valid responses. This process yielded eleven datasets in total: the gold standard dataset (GS0), which consisted exclusively of responses that met the validity criteria, and datasets D1-D10, which contained between 10% and 100% of reportedly invalid responses from respondents who failed the attention check (Figure 2). This controlled change of the survey data reflects empirically realistic conditions in which invalid responses are intermingled with ostensibly valid ones, remaining undetected at the time of analysis. Our experimental design provides a sensitivity analysis of whether and at what point supposed changes in survey data quality affect persona generation.

3.3 Persona Generation Approach

We generated personas from all eleven datasets using Survey2Persona (S2P) [23], a publicly available APG system. S2P is designed to transform survey responses into structured persona profiles. S2P supports multiple survey question types, including Likert-scale, binary, categorical, open-ended, and demographic data, which enables researchers and developers to construct personas directly from survey datasets. Via its online workflow, end users select one or more survey items as the cognitive lens for S2P to process using clustering, an established method in data-driven persona research [42], to identify patterns and produce representative persona archetypes.

S2P uses demographic variables (e.g., age, gender, nationality) along with a substantially larger set of behavioral and attitudinal survey items to generate personas, thereby contextualizing clusters derived primarily from behavioral patterns. The S2P process uses the full set of survey variables (115), reducing the likelihood that demographic variables will overly dominate persona creation. The number of personas (16 in this research) aligns with prior quantitative persona research, which recommends larger persona sets for heterogeneous populations [14]. Importantly, our experimental objective is to examine whether commonly used internal survey validity checks, specifically attention checks, produce detectable changes in downstream research artifacts. If these validity checks can distinguish low-quality from high-quality responses, then the proportion of responses that fail these checks would be expected to produce observable changes in persona outputs.

We used an identical persona-generation procedure with S2P across all datasets (GS0, D01–D10) to ensure that any differences in persona were due solely to changes in the survey data. By holding the persona generation process constant, the design isolates survey data as the experimental variable to examine whether changes

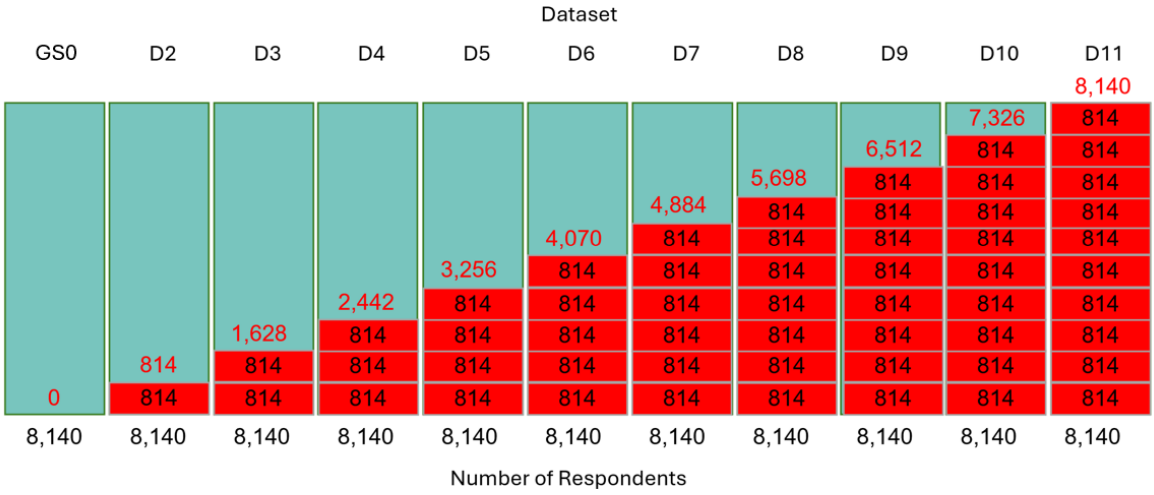


Figure 2: Process of building test datasets from the baseline dataset. From the 8,140 good responses (green) baseline, we progressively replaced valid responses with invalid responses in increments of 814 (red).

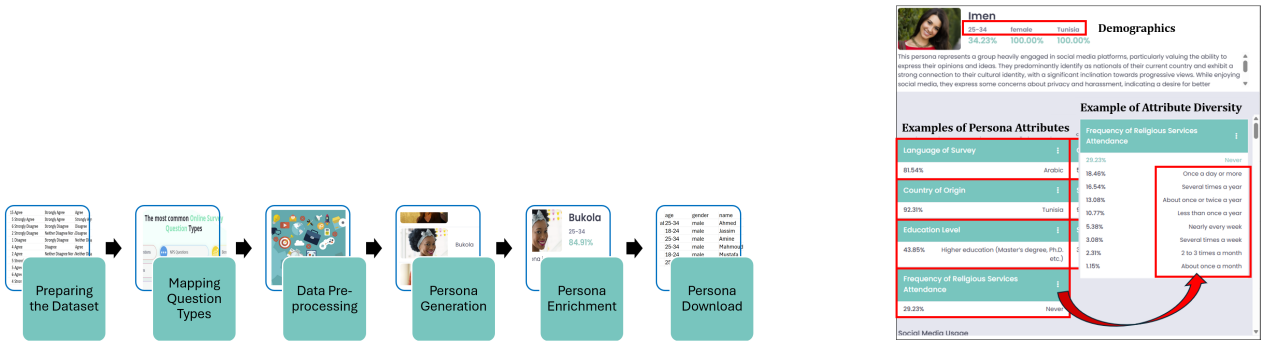


Figure 3: (a) The persona creation approach of Survey2Persona. The core algorithmic method is hierarchical clustering, the most common approach for creating quantitative personas. (b) A persona profile generated from S2P showing the three levels of analysis: (a) demographic variables, (b) the persona profile attributes, and (c) the persona attribute value diversity.

produce detectable effects at the persona level. As S2P is publicly accessible and the data are released as supplementary material, the study is replicable, supporting transparency in research and enabling independent verification [44].

The persona generation workflow applied across all eleven datasets is illustrated in Figure 3(a). First, we formatted the survey responses as CSV files and loaded them into S2P for processing. We then mapped survey items to corresponding persona profile fields to preserve alignment between survey constructs and persona attributes. We performed data preprocessing, including data cleaning, handling missing values, and converting the data into compatible formats. We generate hierarchical clustering, a common technique in persona research [38]. We employed three input variables for persona construction: (a) demographic attributes of age, gender, and nationality, (b) behavioral variables describing platform usage patterns, and (c) attitudinal or preference variables derived from Likert-scale survey items. This approach aligns with methodologies of identifying patterns of user behavior and preferences [14], with the demographics humanizing the data. S2P then groups and summarizes the survey items, producing persona profiles with descriptive headlines, narratives, and standard persona attributes.

We generated sixteen personas for each dataset. Sixteen is consistent with recommendations for large-scale datasets, as a larger persona set improves representation of heterogeneous populations [46] and supports more inclusive persona-based design practices [14]. Sixteen personas provide greater granularity in representing heterogeneous user populations [46] and mitigate limitations associated with smaller persona sets [47]. Our use of a large persona set strengthens the study's ability to detect potential effects of survey data degradation. If changes in data quality meaningfully influenced persona outputs, such effects would be more likely to surface across a richer set of profiles. To address potential ordering effects, we observed no differences in persona generation. Nonetheless, we

randomly shuffled responses across all datasets to ensure an even distribution across conditions. We executed each experimental condition three times on S2P, producing identical persona sets across iterations, confirming the stability of the algorithmic process itself.

3.4 Evaluation Measurement

We conducted multilevel comparative analysis across the persona sets from each dataset. First, we examined demographic overlap by comparing distributions of gender, age, and nationality across persona sets. Second, we analyzed attribute overlap by identifying persona attributes shared across sets [30]. Although some prior research suggests that a small number of core attributes typically drive most interpretation and design, we believe that analyzing all attributes preserves the breadth of persona attributes. Finally, we operationalized the actual diversity of survey values [18]. to enable an evaluation of whether changes in the survey data [20] increase variability in persona representations, as the survey values underpin the attributes in the persona profiles. Each of these levels of analysis is illustrated in Figure 3(b).

We employed the Overlap Coefficient (a.k.a., Szymkiewicz–Simpson coefficient) [54] to compare demographic attributes and persona profile items across persona sets, as it is well-suited for assessing whether changes in survey data quality produce detectable differences in persona outputs. The Overlap Coefficient (OC) quantifies similarity between two finite sets by relating the size of their intersection to the size of the smaller set. Formally, for two sets A and B, it is defined as: $\text{Overlap}(A, B) = \frac{|A \cap B|}{\min(|A|,|B|)}$. Because the denominator is the smaller of the two sets, if the sets are of different sizes, the coefficient ranges from 0 to 1, where 1 indicates complete containment of the smaller set within the larger or identical set, and 0 indicates no shared elements. The OC is appropriate for our study, as it allows comparisons across persona sets and provides a measure for detecting whether variations in survey data translate into differences among personas. As OC is a measure of similarity, 1-OC is a measure of difference. We then employed ANOVA for comparisons across sets. ANOVA is a conventional approach for comparing group means across multiple experimental conditions in HCI and persona research [13], and for comparisons involved a small, fixed number of conditions (GS0 versus D1–D10) with comparable sample sizes across sets, satisfying the test's structural requirements. Also, ANOVA is quite robust to non-normalized data [19].

4 Results

Contrary to our Hs, introducing reportedly invalid data into supposedly valid data did not result in notable changes in the generated personas. Persona sets remained essentially unchanged across datasets, regardless of the level of invalid data.

4.1 Demographic Characteristics

As shown in Table 1, the introduction of invalid survey data to the supposedly valid data did not yield a systematic change in persona demographics. A reminder that the demographic survey questions were split: one at the beginning and two towards the end, to mitigate the risk of response fatigue affecting/not affecting demographics. Across six of the ten experimental comparisons, the persona demographics of the sixteen personas in each set remained unchanged relative to the gold standard set GS0. In the remaining four comparisons, only a single persona differed from the sixteen, and in each instance, the change was limited to an age category, with no variation observed in other demographic attributes. For the six comparisons exhibiting identical values, we could not perform an ANOVA without producing a degenerate p-value (i.e., the sets were identical). In the four remaining comparisons, one-way ANOVA revealed no statistically significant differences between the gold standard and the experimental datasets ($F(2, 30) = 0$, $p = 1$). Accordingly, **H01 is not supported**: *increases in invalid survey data did not produce corresponding increases in demographic instability.*

4.2 Persona Profile Attributes

Table 2 presents the results of the analysis of persona profile item consistency. Unlike demographic attributes, the changes in the survey data did produce some changes in persona profile items across datasets. As shown in Table 2, differences emerged between persona sets as the proportion of invalid data increased. However, the magnitude of these shifts remained modest, suggesting that differences in the underlying datasets persisted. The resulting changes in persona profiles are limited, with about 1 attribute in 10 changed from the gold-standard baseline. GS0.

As shown in Table 2, we observed statistically significant differences between each experimental condition (D1–D10) and the baseline dataset (D1) when comparing the values of the persona profile items. Accordingly, **H02 is supported**: *increases in the proportion of invalid survey data are associated with increased changes in persona profile attributes.* However, readers should interpret this finding cautiously. Although the effect is statistically detectable, the changes are practically limited across all persona attributes. While persona profile values exhibit sensitivity to changes in survey data, the magnitudes do not provide a reliable basis for determining whether the underlying survey data is valid. Of the 115 possible persona attributes in a persona profile, about 14 items (12.2%) changed per persona set, with the change nearly constant across the ten datasets, regardless of the percentage of invalid survey data.

4.3 Survey Value Diversity

As shown in Table 3, changes in the survey data did not significantly affect the diversity of persona attribute values. Across datasets, changes in the range of attribute values were minimal: two datasets exhibited a single novel response, two datasets exhibited two novel responses, and six datasets exhibited three novel responses. These did not result in statistically significant differences when compared to the gold standard dataset, GS0, as confirmed by one-way ANOVA ($F(1, 196) = 0$, $p > 0.99$). Accordingly, **H03 is not supported**; *survey value diversity does not increase as the proportion of invalid survey data increases.*

4.4 Demographic Characteristics Using the K-prototypes Clustering Algorithm

We also conducted an assessment to determine whether the results would be similar or different when using another persona-generation algorithm (though the analysis for RQ3 was independent

Table 1: Overlap results for demographics (similarity and difference) compared to the GS0 dataset. Datasets with identical overlap are shaded.

	Dataset									
	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
OC	0.94	0.94	1.00	0.94	1.00	1.00	1.00	1.00	1.00	0.94
1-OC	0.06	0.06	0.00	0.06	0.00	0.00	0.00	0.00	0.00	0.06

Table 2: Results for attribute (similarity and difference) compared to the GS0 dataset.

	Dataset									
	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
OC	0.86	0.89	0.89	0.86	0.89	0.88	0.87	0.89	0.84	0.89
1-OC	0.14	0.11	0.11	0.14	0.11	0.12	0.13	0.11	0.16	0.11

Table 3: Results for survey value diversity (similarity and difference) compared to the GS0 dataset.

	Dataset									
	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
OC	0.998	0.998	0.997	0.997	0.995	0.995	0.995	0.995	0.995	0.995
1-OC	0.002	0.002	0.003	0.003	0.005	0.005	0.005	0.005	0.005	0.005

of any persona creation algorithm). We selected the k-prototypes clustering algorithm, as it is a common approach in persona generation. For this check, we compared the GS0 gold standard dataset to D10, the dataset with all invalid data for the demographic overlap. We used all the survey questions and set the number of personas to 15. Comparing demographics, the OC was 0.40 (i.e., 6 of the 15 personas were the same), lower than with hierarchical clustering but still quite high. Also, the remaining personas typically varied by only one demographic attribute (generally age or nationality). This finding shows that, as one would expect, the algorithmic choice does influence the number of overlapping personas, but overlapping personas still exist, and, with at least these two algorithms, it is quite high.

5 Discussion and Implications

These findings address the critical issue of why the changes in the survey data produce so little change in personas. H1 (Changes in persona demographics) was not supported. Persona demographics are not responsive. H2 (Changes in persona profile attributes) was partially supported. Persona attributes are somewhat responsive. H3 (Diversity in attributes) was not supported. Attribute value ranges are generally unchanged. Findings show that attention check credibility decreases as the sample size for the survey increases because the probability of a response being selected increases for large numbers of respondents, rendering the distributions of valid and invalid respondents indistinguishable at scale [53]. At smaller sample sizes, internal checks such as attention checks may be meaningful. However, given the scale common in HCI and modern survey-based research, internal checks alone become insufficient for establishing survey data validity [39].

5.1 Theoretical Implications

Large samples are often used to capture the segments within heterogeneous user populations [26]. However, our findings show that as the survey sample size increases, the number of plausible response combinations of bounded-choice items (e.g., Likert and binary items) grows, increasing the likelihood that invalid responses become statistically indistinguishable from valid responses [10]. This study demonstrates that the validity of survey data cannot be reliably inferred from the survey data itself or from artifacts derived from it, such as personas. The findings challenge a straightforward application of the GIGO assumption. With survey data, relying solely on internal measures such as attention checks does not allow one to determine what is "garbage" and what is not. Based on our findings, large-scale surveys (i.e., surveys with many respondents and questionnaire items) can mask both valid and invalid responses, producing personas that appear stable yet are grounded in unreliable data—*a paradox of data stability*. Therefore, survey-based personas and related practices are placed into question, as using survey data may create internally coherent personas that still misrepresent users. Importantly, these results suggest that commonly used internal validity controls (e.g., attention checks, completion time thresholds) are insufficient, particularly for lengthy surveys administered at scale. Large survey samples make it practically impossible to determine valid from invalid responses.

Lengthy survey instruments may encourage satisficing or careless responding, while large sample sizes amplify response heterogeneity, further obscuring the presence of invalid data (i.e., with more respondents, the probability increases that each response to a question is chosen). Survey data (e.g., Boolean, Likert, and Multiple-Choice items) can obscure the validity of responses due to the constrained nature of survey response formats. We refer

to this as the Validity Masking Effect. Employing finite response categories (Boolean, Likert, multiple-choice) means that any response is already plausible. Given this pigeonhole principle setup, the Validity Masking Effect states *determining validity is practically impossible from the survey data itself for finite response categories, as the probability is that all options are selected by someone as the sample size grows.* Many survey items have a finite set of predefined options, so large-scale surveys inherently exhibit significant response variance. For instance, binary questions force all responses into "yes" or "no" categories, producing seemingly stable patterns regardless of the respondent's intent or accuracy. Likert scales allow respondents to choose from a handful of options, rendering responses statistically indistinguishable in large samples where all options are likely to be selected by at least some respondents. These structural constraints make it impossible to distinguish valid from invalid survey responses using internal survey methods, such as attention checks, and, by extension, to distinguish between valid and invalid personas derived from that data. The Validity Masking Effect is a critical limitation of relying on survey data for research, which often includes Boolean, multiple-choice, or Likert-scale questions. These factors increase the difficulty of determining survey data validity and point to the need for alternative approaches, such as shorter, more targeted surveys with constrained sampling frames, and the incorporation of validation techniques derived from established methods in computer science and statistical analysis.

5.2 Practical Implications

Our findings challenge the assumption that attention checks (and perhaps other internal survey measures) ensure the collection of higher-quality survey data. From our large-scale survey, responses from participants who passed the attention checks were nearly indistinguishable from those who failed, indicating that such internal measures have a limited effect on survey data validity. This finding is supported by some prior research indicating that attention checks are commonly used [39], but their effectiveness is questionable [51]. Attention checks may create a false sense of methodological rigor while masking broader reliability problems in survey data used for persona generation and other uses. Failing attention checks have been used to flag invalidity [33, 36]. However, our findings show that responses from those failing attention checks were mostly statistically indistinguishable from those of those passing attention checks [52]. Therefore, attention checks are not a means for certifying validity, and, at least for large scale surveys, for ensuring validity.

Now what to do about it? First, this finding requires the HCI research community to reconsider lengthy surveys in favor of shorter, targeted surveys that make data validation more manageable, especially for tasks such as persona creation. Shorter surveys have the advantage of avoiding respondent fatigue and a greater focus on a limited number of research constructs. Second, findings point to the need to cap the number of respondents per survey, as response variability with bounded choice questions increases with the number of respondents, making it more difficult to assess data validity (the Validity Masking Effect). In situations where large survey samples are needed, HCI and other researchers need to leverage validation mechanisms, such as stratified data-quality audits, response-pattern analysis, or qualitative verification. However, even these methods need to be evaluated, along with pursuing other novel approaches [6]. With personas created from survey data, independent data sources are needed for verification once the personas have been created.

5.3 Strengths, Limitations, and Future Research

The strength of this research for the community is that it systematically introduced invalid survey responses into valid data, enabling a controlled examination of how survey data degradation affects persona generation and exposing the practical difficulty of detecting invalid survey data. While the large-scale dataset is representative of many research surveys, future studies should investigate whether the invalidatable effect also occurs in smaller surveys or in surveys conducted with smaller samples. It would be interesting to determine whether the effectiveness of attention checks and other internal mechanisms is inversely proportional to the sample size, for example. Although the professional survey provider implemented the attention checks employed as part of their standard data-quality protocol, future research could investigate the effect of other types of attention checks and different validation mechanisms. Finally, an exhaustive investigation of algorithmic selection for persona creation would be quite beneficial for determining which algorithms are most responsive to minuscule changes in survey responses and which are not.

6 Conclusion

From the premise that survey data is invalidatable, this research highlights that survey data validity is difficult, if not impossible, to determine solely through internal means, such as attention checks. As such, any downstream artifacts, such as personas, created from this survey data mask data quality issues, and these artifacts are themselves difficult to distinguish as valid or invalid. The implications are that relying solely on survey data in HCI research is problematic and is fraught with risks due to invalidatable issues.

References

[1] James D. Abbey and Margaret G. Meloy. 2017. Attention by design: Using attention checks to detect inattentive respondents and improve data quality. *Journal of Operations Management* 53, (2017), 63–70.

[2] J. An, H. Kwak, S. Jung, J. Salminen, M. Admad, and B. Jansen. 2018. Imaginary People Representing Real Numbers: Generating Personas from Online Social Media Data. *ACM Trans. Web* 12, 4 (2018), 1–26. https://doi.org/10.1145/3265986

[3] David La Barbera, Eddy Maddalena, Michael Soprano, Kevin Roitero, Gianluca Demartini, Davide Ceolin, Damiano Spina, and Stefano Mizzaro. 2024. Crowdsourced Fact-checking: Does It Actually Work? *Information Processing & Management* 61, 5 (2024), 103792. https://doi.org/10.1016/j.ipm.2024.103792

[4] Elisa Bertino, Chenyun Dai, and Murat Kantarcioglu. 2009. The Challenge of Assuring Data Trustworthiness. In *Database Systems for Advanced Applications*, Xiaofang Zhou, Haruo Yokota, Ke Deng and Qing Liu (eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 22–33. https://doi.org/10.1007/978-3-642-00887-0_2

[5] Paul P. Biemer. 2020. Data Quality and Inference Errors. In *Big Data and Social Science* (2nd ed.). Chapman and Hall/CRC.

[6] Akashdeep Chakraborty and Joseph J. LaViola Jr. 2026. From Narrative to Numbers: Evaluating Survey Questionnaires with Large Language Models. In *Proceedings of the 31st International Conference on Intelligent User Interfaces*, March 23, 2026. ACM, Paphos Cyprus, 1647–1657. https://doi.org/10.1145/3742413.3789071

[7] Christopher N. Chapman, Edwin A. Love, Russell P. Milham, Paul ElRif, and James L. Alford. 2008. Quantitative Evaluation of Personas as Information. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 52, (2008), 1107–1111. https://doi.org/10.1177/154193120805201602

[8] D. Chicco, L. Oneto, and E. Tavazzi. 2022. Eleven quick tips for data cleaning and feature engineering. *PLOS Computational Biology* 18, (2022). https://doi.org/10.1371/journal.pcbi.1010718

[9] Alan Cooper. 2004. The Inmates Are Running the Asylum: Why High Tech Products Drive Us Crazy and How to Restore the Sanity (2nd ed.). Pearson Higher Education.

[10] Paul G. Curran. 2016. Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology* 66, (2016), 4–19.

[11] Jessica Daikeler, Leon Fröhling, Indira Sen, Lukas Birkenmaier, Tobias Gummer, Jan Schwalbach, Henning Silber, Bernd Weiß, Katrin Weller, and Clemens Lechner. 2024. Assessing Data Quality in the Age of Digital Social Research: A Systematic Review. *Social Science Computer Review* (April 2024), 08944393241245395. https://doi.org/10.1177/08944393241245395

[12] Sven Eckhardt, Merlin Knaeble, Andreas Bucher, Dario Staehelin, Mateusz Dolata, Doris Agotai, and Gerhard Schwabe. 2023. "Garbage In, Garbage Out": Mitigating Human Biases in Data Entry by Means of Artificial Intelligence. In *Human-Computer Interaction – INTERACT 2023*, José Abdelnour Nocera, Marta Kristín Lárusdóttir, Helen Petrie, Antonio Piccinno and Marco Winckler (eds.). Springer Nature Switzerland, Cham, 27–48. https://doi.org/10.1007/978-3-031-42286-7_2

[13] Andy Field. 2024. *Discovering statistics using IBM SPSS statistics*. Sage publications limited. Retrieved July 13, 2026 from https://books.google.com/books?hl$=$en&lr$=$&id$=$83L2EAAAQBAJ&oi$=$fnd&pg$=$PA18&dq$=$Discovering+statistics+using+IBM+SPSS+statistics&ots$=$UbQUxlGPHP&sig$=$SVhTamPXR8fUpWwFGIEWJdKCADY

[14] Joy Goodman-Deane, Sam Waller, Dana Demin, Arantxa González-de-Heredia, Mike Bradley, and John P. Clarkson. 2018. Evaluating inclusivity using quantitative personas. (2018). Retrieved August 27, 2024 from https://dl.designresearchsociety.org/drs-conference-papers/drs2018/researchpapers/133/

[15] Kathleen W. Guan, Joni Salminen, Soon-Gyo Jung, and Bernard J. Jansen. 2023. Leveraging Personas for Social Impact: A Review of Their Applications to Social Good in Design. *International Journal of Human–Computer Interaction* (September 2023), 1–16. https://doi.org/10.1080/10447318.2023.2247568

[16] Tom Haegemans, Monique Snoeck, and Wilfried Lemahieu. 2019. A theoretical framework to improve the quality of manually acquired data. *Information & Management* 56, 1 (January 2019), 1–14. https://doi.org/10.1016/j.im.2018.05.014

[17] Arantxa Gonzalez de Heredia, Joy Goodman-Deane, S. Waller, P. J. Clarkson, D. Justel, I. Iriarte, and Jesús Hernández. 2018. Personas for Policy Making and Healthcare Design. (2018), 2645–2656. https://doi.org/10.21278/IDC.2018.0438

[18] Todd J. Hostager and Kenneth P. De Meuse. 2002. Assessing the Complexity of Diversity Perceptions: Breadth, Depth, and Balance. *Journal of Business and Psychology* 17, 2 (December 2002), 189–206. https://doi.org/10.1023/A:1019681314837

[19] David Hull. 1993. Using statistical testing in the evaluation of retrieval experiments. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '93*, 1993. ACM Press, Pittsburgh, Pennsylvania, United States, 329–338. https://doi.org/10.1145/160688.160758

[20] Bernard J. Jansen, Soon-gyo Jung, and Joni Salminen. 2023. Employing large language models in survey research. *Natural Language Processing Journal* 4, (2023), 100020.

[21] Bernard J. Jansen, Joni Salminen, Soon-gyo Jung, and Hind Almerekhi. 2022. The illusion of data validity: Why numbers about people are likely wrong. *Data and Information Management* 6, 4 (2022), 100020. https://doi.org/10.1016/j.dim.2022.100020

[22] Najeeb Moharram Jebreel, Rami Haffar, Ashneet Khandpur Singh, David Sánchez, Josep Domingo-Ferrer, and Alberto Blanco-Justicia. 2020. Detecting Bad Answers in Survey Data Through Unsupervised Machine Learning. In *Privacy in Statistical Databases*, 2020. Springer International Publishing, Cham, 309–320. https://doi.org/10.1007/978-3-030-57521-2_22

[23] Soon-Gyo Jung, Joni Salminen, Kholoud Khalil Aldous, and Bernard J. Jansen. 2025. PersonaCraft: Leveraging language models for data-driven persona development. *International Journal of Human-Computer Studies* 197, (2025), 103445.

[24] Carolin Kaiser, Jakob Kaiser, Vladimir Manewitsch, Lea Rau, and Rene Schallner. 2025. Simulating Human Opinions with Large Language Models: Opportunities and Challenges for Personalized Survey Data Modeling. In *Adjunct Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization*, June 16, 2025. ACM, New York City USA, 82–86. https://doi.org/10.1145/3708319.3733685

[25] Melissa G. Keith and Alexander S. McKay. 2024. Too Anecdotal to Be True? Mechanical Turk Is Not All Bots and Bad Data: Response to Webb and Tangney (2022). *Perspect Psychol Sci* (March 2024), 17456916241234328. https://doi.org/10.1177/17456916241234328

[26] Jon A. Krosnick. 1991. Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology* 5, 3 (May 1991), 213–236. https://doi.org/10.1002/acp.2350050305

[27] Jiwei Li, Michel Galley, Chris Brockett, Georgios P. Spithourakis, Jianfeng Gao, and W. Dolan. 2016. A Persona-Based Neural Conversation Model. *ArXiv* abs/1603.06155, (2016). https://doi.org/10.18653/v1/P16-1094

[28] Catherine C. Marshall and Frank M. Shipman. 2013. Experiences surveying the crowd: reflections on methods, participation, and reliability. In *Proceedings of the 5th Annual ACM Web Science Conference* (*WebSci '13*), 2013. Association for Computing Machinery, New York, NY, USA, 234–243. https://doi.org/10.1145/2464464.2464485

[29] Tara Matthews, Tejinder Judge, and Steve Whittaker. 2012. How do designers and user experience professionals actually perceive and use personas? In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems - CHI '12*, 2012. ACM Press, Austin, Texas, USA, 1219. https://doi.org/10.1145/2207676.2208573

[30] Tara Matthews, Tejinder Judge, and Steve Whittaker. 2012. How do designers and user experience professionals actually perceive and use personas? In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, May 05, 2012. ACM, Austin Texas USA, 1219–1228. https://doi.org/10.1145/2207676.2208573

[31] Jennifer (Jen) McGinn and Nalini Kotamraju. 2008. Data-driven persona development. In *Proceeding of the twenty-sixth annual CHI conference on Human factors in computing systems - CHI '08*, 2008. ACM Press, Florence, Italy, 1521. https://doi.org/10.1145/1357054.1357292

[32] A. Meade and S. Craig. 2012. Identifying careless responses in survey data. *Psychological methods* 17 3, (2012), 437–55. https://doi.org/10.1037/a0028085

[33] Adam W. Meade and S. Bartholomew Craig. 2012. Identifying careless responses in survey data. *Psychological methods* 17, 3 (2012), 437.

[34] Tomasz Miaskiewicz and Kenneth A. Kozar. 2011. Personas and user-centered design: How can personas benefit product design processes? *Design studies* 32, 5 (2011), 417–430.

[35] Donna L. Mitchell, Gary Klein, and Joseph L. Balloun. 1996. Mode and gender effects on survey data quality. *Information & Management* 30, 1 (January 1996), 27–34. https://doi.org/10.1016/0378-7206(95)00038-0

[36] Daniel M. Oppenheimer, Tom Meyvis, and Nicolas Davidenko. 2009. Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of experimental social psychology* 45, 4 (2009), 867–872.

[37] C. Putnam, B. Kolko, and S. Wood. 2012. Communicating about users in ICTD: leveraging HCI personas. *Proceedings of the Fifth International Conference on Information and Communication Technologies and Development* (2012). https://doi.org/10.1145/2160673.2160714

[38] Sinziana I. Rasca, Karin Markvica, and Benjamin Biesinger. 2023. Persona Design Methodology for Work-Commute Travel Behaviour Using Latent Class Cluster Analysis. *Multimodal Transportation* 2, 4 (December 2023), 100095. https://doi.org/10.1016/j.multra.2023.100095

[39] Yefim Roth and Ofir Yakobi. 2024. Attention! Do We Really Need Attention Checks? *Behavioral Decision Making* 37, 2 (April 2024), e2377. https://doi.org/10.1002/bdm.2377

[40] R. Rust and B. Cooil. 1994. Reliability Measures for Qualitative Data: Theory and Implications. *Journal of Marketing Research* 31, (1994), 1–14. https://doi.org/10.1177/002224379403100101

[41] Joni Salminen, Kamal Chhirang, Soon-Gyo Jung, Saravanan Thirumuruganathan, Kathleen W. Guan, and Bernard J. Jansen. 2022. Big Data, Small Personas: How Algorithms Shape the Demographic Representation of Data-Driven User Segments. *Big Data* 10, 4 (August 2022), 313–336. https://doi.org/10.1089/big.2021.0177

[42] Joni Salminen, Kathleen Guan, Soon-Gyo Jung, and Bernard J. Jansen. 2021. A Survey of 15 Years of Data-Driven Persona Development. *International Journal of Human–Computer Interaction* 37, 18 (2021), 1685–1708. https://doi.org/10.1080/10447318.2021.1908670

[43] Joni Salminen, Soon-gyo Jung, Shammur Absar Chowdhury, Sercan Sengün, and Bernard J Jansen. 2020. Personas and Analytics: A Comparative User Study of Efficiency and Effectiveness for a User Identification Task. In *Proceedings of the ACM Conference of Human Factors in Computing Systems (CHI'20)*, 2020. ACM, Honolulu, Hawaii, USA. https://doi.org/https://doi.org/10.1145/3313831.3376770

[44] Joni Salminen, Soon-Gyo Jung, and Bernard Jansen. 2022. Developing Persona Analytics Towards Persona Science. In *27th International Conference on Intelligent User Interfaces* (*IUI '22*), 2022. Association for Computing Machinery, New York, NY, USA, 323–344. https://doi.org/10.1145/3490099.3511144

[45] Joni Salminen, Soon-Gyo Jung, and Bernard Jansen. 2022. Developing Persona Analytics Towards Persona Science. In *27th International Conference on Intelligent User Interfaces* (*IUI '22*), March 22, 2022. Association for Computing Machinery, New York, NY, USA, 323–344. https://doi.org/10.1145/3490099.3511144

[46] Joni Salminen, Soon-Gyo Jung, Lene Nielsen, and Bernard Jansen. 2022. Creating More Personas Improves Representation of Demographically Diverse Populations: Implications Towards Interactive Persona Systems. In *Nordic Human-Computer Interaction Conference*, 2022. ACM, Aarhus Denmark, 1–11. https://doi.org/10.1145/3546155.3546654

[47] Joni Salminen, Soon-gyo Jung, Lene Nielsen, Sercan Şengün, and Bernard J. Jansen. 2022. How does varying the number of personas affect user perceptions and behavior? Challenging the 'small personas' hypothesis! *International Journal of Human-Computer Studies* 168, (2022), 102915.

[48] Joni O. Salminen, Kathleen W. Guan, Soon-Gyo Jung, S. A. Chowdhury, and B. Jansen. 2020. A Literature Review of Quantitative Persona Creation. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (2020). https://doi.org/10.1145/3313831.3376502

[49] Joni Salminen, Joao M. Santos, Haewoon Kwak, Jisun An, Soon-gyo Jung, and Bernard J. Jansen. 2020. Persona perception scale: development and exploratory validation of an instrument for evaluating individuals' perceptions of personas. *International Journal of Human-Computer Studies* 141, (2020), 102437.

[50] David A. Siegel. 2010. The mystique of numbers: belief in quantitative approaches to segmentation and persona development. In *CHI '10 Extended Abstracts on Human Factors in Computing Systems*, April 10, 2010. ACM, Atlanta Georgia USA, 4721–4732. https://doi.org/10.1145/1753846.1754221

[51] Henning Silber, Joss Roßmann, and Tobias Gummer. 2022. The Issue of Noncompliance in Attention Check Questions: False Positives in Instructed Response Items. *Field Methods* 34, 4 (November 2022), 346–360. https://doi.org/10.1177/1525822X221115830

[52] Henning Silber, Joss Roßmann, and Tobias Gummer. 2022. The Issue of Noncompliance in Attention Check Questions: False Positives in Instructed Response Items. *Field Methods* 34, 4 (November 2022), 346–360. https://doi.org/10.1177/1525822X221115830

[53] David L. Vannette and Jon A. Krosnick. 2014. Answering Questions: *A Comparison of Survey Satisficing and Mindlessness*. In *The Wiley Blackwell Handbook of Mindfulness* (1st ed.), Amanda Ie, Christelle T. Ngnoumen and Ellen J. Langer (eds.). Wiley, 312–327. https://doi.org/10.1002/9781118294895.ch17

[54] M. K. Vijaymeena and K. Kavitha. 2016. A survey on similarity measures in text mining. *Machine Learning and Applications: An International Journal* 3, 2 (2016), 19–28.

[55] Le-Hung Vu and K. Aberer. 2007. A Probabilistic Framework for Decentralized Management of Trust and Quality. (2007), 328–342. https://doi.org/10.1007/978-3-540-75119-9_23

[56] Steven Euijong Whang, Yuji Roh, Hwanjun Song, and Jae-Gil Lee. 2021. Data collection and quality challenges in deep learning: a data-centric AI perspective. *The VLDB Journal* (2021), 1–23. https://doi.org/10.1007/s00778-022-00775-9

[57] Runting Zhong, Saihong Han, and Zi Wang. 2024. Developing personas for live streaming commerce platforms with user survey data. *Univ Access Inf Soc* 23, 4 (November 2024), 1705–1721. https://doi.org/10.1007/s10209-023-00996-x